Sairaanhoitoalueen kokonaiskustannusten ja riskitapausten mallintaminen

Lauri Pyyhkälä
Helsingin yliopisto
Matematiikan ja tilastotieteen osasto
Pro Gradu -tutkielma

Syksy 2020

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Matemaattis-luonnontieteellinen	Matematiikan ja tilastotieteen maisteriohjelma

Tekijä — Författare — Author
Lauri Pyyhkälä

Työn nimi — Arbetets titel — Title

Sairaanhoitoalueen kokonaiskustannusten ja riskitapausten mallintaminen

Oppiaine — Läroämne — Subject
Matematiikka

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Syksy 2020	66 s.

Tiivistelmä — Referat — Abstract

Työssä tutkitaan Hyvinkään sairaanhoitoalueen kustannuksia, sekä kokonaiskustannusten tasolla, että yksittäisen potilaan tasolla. Sairaanhoidon kustannukset ovat olennainen osa yhteiskunnan toimintaa ja ne vaikuttavat merkittävästi kuntien ja kaupunkien talouteen. Tämän takia on hyödyllistä pystyä ymmärtämään ja mallintamaan näitä kustannuksia. Aineistona on käytetty HUSilta saatua dataa kustannuslajeista, potilaista ja diagnoosiryhmistä. Tutkimuksen ensimmäinen tavoite on löytää tilastollinen malli, jolla voidaan ennustaa kokonaiskustannuksia. Toisena tavoitteena on löytää yksittäisten potilaiden kustannuksiin sopiva jakauma.

Työn alussa esitellään todennäköisyysteoriaa ja tilastollisia menetelmiä, joita hyödynnetään tutkimuksessa. Näistä tärkeimmät ovat keskineliövirhe, aikasarjamalli ja tilastolliset testit. Näiden teorioiden avulla luodaan mallit kokonaiskustannuksille ja yksittäisen potilaan kustannuksille.

Kokonaiskustannusten analysointi aloitetaan erottelemalla suurimmat kustannuslajit, jotta niiden tutkiminen olisi selkeämpää. Näihin isoimpiin kustannuslajeihin valitaan tärkeimmät selittävät muuttujat käyttämällä lineaarista regressiomallia ja informaatiokriteeriä. Näin saatujen muuttujien avulla voidaan muodostaa moniulotteinen aikasarjamalli kokonaiskustannuksille ja tärkeimmille muuttujille. Tämän mallin avulla voidaan luoda ennuste tulevaisuuden kustannuksista, kun se on validoitu muun aineiston avulla.

Tutkielman viimeisessä osiossa tutustutaan tarkemmin paksuhäntäisiin jakaumiin, ja esitellään niiden tärkeimpiä ominaisuuksia. Paksuhäntäisillä jakaumilla suurien havaintojen todennäköisyys on merkittävästi suurempi kuin kevythäntäisillä. Tämä vuoksi niiden tunnistaminen on tärkeää, sillä paksuhäntäiset jakaumat voivat aiheuttaa merkittäviä kustannuksia. Termien esittelyn jälkeen tehdään visuaalista tarkastelua potilaiden kustannuksista. Tavoitteena on selvittää, mikä jakauma kuvaisi parhaiten potilaiden kustannuksia. Tutkimuksessa verrataan erilaisten teoreettisten jakaumien kuvaajia aineistosta laskettuun empiiriseen jakaumaan. Erilaisista kuvaajista voidaan päätellä, että kustannusten jakauma on paksuhäntäinen. Lisäksi huomataan, että havainnot sopisivat yhteen sen oletuksen kanssa, että jakauman häntä muistuttaa ainakin asymptoottisesti potenssihäntää.

Työn lopussa perustellaan ääriarvoteoriaan nojaten, miksi potenssihännät ovat luonnollinen malli suurimmille kustannuksille.

Avainsanat — Nyckelord — Keywords
Tilastotiede, Aikasarja, Riskiteoria, Paksuhäntäinen jakauma

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpulan tiedekirjasto

Muita tietoja — Övriga uppgifter — Additional information

Sisällys

1 Johdanto

Tutkimus käsittelee Helsingin ja Uudenmaan sairaanhoitopiirin toiminnan kustannuksia ja erityisesti Hyvinkään sairaanhoitoaluetta. Tutkimuksen perimmäisenä tarkoituksena on pyrkiä ennustamaan sairaanhoitopiirin kustannuksia tulevaisuudessa sekä kokonaiskustannusten tasolla että yksittäisen potilaan tasolla. HUS kattaa koko eteläisen Suomen erikoissairaanhoidon sekä joidenkin erikoisalojen hoidon koko Suomen alueelta. HUSin varsinaisen toimialueen väestömäärä on yli puolitoista miljoonaa, ja HUS on suurin terveydenhuollon toimija Suomessa. HUSin kokonaiskulut ovat noin 2 miljardia euroa vuodessa, joten kokonaisuudessaan HUSin toiminnassa liikkuu huomattavat määrät rahaa. HUSin rahoitus on myös osa poliittista keskustelua ja päätöksentekoa, sillä HUS saa julkisena toimijana rahoituksensa pääosin kunnilta.

Tutkimuksen motiivina on pystyä paremmin ennustamaan kustannusten muutokset tulevaisuudessa ja hahmottamaan mitkä tekijät vaikuttavat kokonaiskustannuksiin. Tässä tutkimuksessa erityistä on se, että asiaa käsitellään sairaanhoitopiirin sisäisestä näkökulmasta, tällöin voidaan tarkastella sairaanhoitopiirin sisäisen toiminnan vaikutusta kustannuksiin. Yleensä ottaen sairaanhoitokuluja tarkastellessa tutkitaan sairaalan ulkopuolisia muuttujia. Näitä muuttujia ovat esimerkiksi väestön ikä ja tulotaso.

Tutkimuksen alussa esitellään matemaattisia peruskäsitteitä todennäköisyyslaskennasta, lineaarisesta mallinnuksesta ja aikasarjoista, joita käytetään varsinaisessa tutkimuksessa. Tavoitteena on muistuttaa lukijaa oleellisista käsitteistä ja termeistä. Itse tutkimuksen alussa esitellään käytössä oleva aineisto ja sen tunnuspiirteitä. Samalla aloitetaan aineiston muokkaaminen mallin rakentamiseen sopivammaksi. Aineistoon sovitetaan lineaarinen malli, jolla pystytään ennustamaan kustannusten kehittymistä. Lineaarisen mallin avulla valikoidaan tärkeimmät kustannuksiin vaikuttavat muuttujat, joita käytetään aikasarjamallinnuksessa.

Tutkimuksessa rajoitutaan Hyvinkään sairaanhoitoalueeseen, sillä se muodostaa noin 10-prosenttia HUSin kokonaiskustannuksista, ja sen toiminta on sopivan monimuotoista. Rajaus Hyvinkääseen poistaa käsittelystä suurimman osan erittäin kalliista potilastapauksista, joista HYKS vastaa, kuten keskoset ja elinsiirrot. Tällaiset tapaukset saattaisivat vaikuttaa kustannuskehitykseen vaikeasti ennustettavalla tavalla, joten kokonaiskustannusten kannalta on parempi lähestyä ongelmaa pienemmällä otoksella.

Hyvinkään sairaanhoitopiirin kokonaiskustannuksista luodaan aikasarjamalli, jossa pyritään muutaman muuttujan avulla ennustamaan tulevaisuuden kustannusten kehitystä. Osana mallintamista tarkastetaan myös mallin toimivuus erilaisten diagnostiikkatyökalujen avulla, joilla voidaan todeta mallin olevan luotettava.

Lopuksi tutkitaan yksittäisten potilaiden kustannusten jakaumaa, jonka todetaan olevan paksuhäntäinen. Tutkimuksessa esitellään joitain paksuhäntäisten jakaumien ominaisuuksia. Lisäksi perustellaan paksuhäntäisen jakauman paksuhäntäisyyden estimointiin käytettävien työkalujen käyttö.

2 Matemaattinen taustateoria

Tässä luvussa esitellään tutkimuksessa tarvittavia käsitteitä ja merkintöjä, sekä oleelliset lauseet joihin tutkimuksen lopputulos perustuu. Ensimmäisenä käsitellään yleistä todennäköisyysteoriaa, johon tilastolliset menetelmät perustuvat. Seuraavaksi esitellään tilastollisen mallin määritelmä ja siihen liittyviä peruskäsitteitä kuten uskottavuusfunktio. Tämän jälkeen esitellään lineaarinen malli, jonka jälkeen käydään lyhyesti läpi aikasarjamalliin liittyviä testejä ja käsitteitä.

2.1 Todennäköisyysteoriaa

Todennäköisyysteoria pohjautuu vahvasti mittateoriaan, joten teorian esittely aloitetaan mittateorian peruskäsitteistä. Yleistä mittateoriaa esitellään tarkemmin monisteen [11] C-osiossa. Sen jälkeen esitellään satunnaismuuttujan käsite. Satunnaismuuttujaan liittyy aina jokin tietty jakauma, joka määrittelee sen käytöksen. Esimerkkejä tunnetuista jakaumista ovat tasa-, normaali- ja eksponenttijakauma. Satunnaismuuttujien teoriaa on esitelty tarkemmin Durrettin kirjassa [5] luvussa 1.

2.1.1 Mittateoriaa

Termien esittely aloitetaan mitallisesta avaruudesta, joka on mitta-avaruutta yksinkertaisempi käsite. Määritellään ensiksi sigma-algebra, joka on mitallisen avaruuden toinen komponentti.

Määritelmä 2.1.1. *Olkoon X joukko ja $\mathcal{P}(X)$ kokoelma sen kaikista osajoukoista. Kokoelma $\mathcal{F} \subseteq \mathcal{P}(X)$ on joukon X sigma-algebra, jos seuraavat ehdot pätevät:*

1. $X \in \mathcal{F}$.
2. *Jos* $A \in \mathcal{F}$, *niin* $A^c \in \mathcal{F}$.
3. *Jos* $A_1, A_2, \ldots \in \mathcal{F}$, *niin* $A_1 \cup A_2 \cup \ldots \in \mathcal{F}$.

Määritelmä 2.1.2. *Pari $(X, \mathcal{F})$ on mitallinen avaruus, jos X on joukko ja $\mathcal{F}$ sen sigma-algebra.*

Mitallisessa avaruudessa on myös määritelty mitalliset joukot.

Määritelmä 2.1.3. *Oletetaan, että on olemassa mitallinen avaruus $(X, \mathcal{F})$. Joukko $A \subseteq X$ on mitallinen, jos $A \in \mathcal{F}$.*

Seuraavaksi esitellään mitan sekä mitta-avaruuden määritelmä, joiden avulla voidaan määritellä todennäköisyysmitta ja -avaruus.

Määritelmä 2.1.4. *Oletetaan, että X on joukko ja $\mathcal{F}$ on sen sigma-algebra. Määritellään, että funktio $m : \mathcal{F} \to [0, \infty]$ on mitta, jos sille pätee seuraavat ehdot:*

1. *Tyhjälle joukolle* $\emptyset$ *pätee* $m(\emptyset) = 0$.

2. *Funktio toteuttaa täysadditiivisuus ehdon eli kaikille erillisille joukoille* $F_i \in \mathcal{F}, i \in \mathbb{N}$ *pätee*

$$m\left(\bigcup_{i\in\mathbb{N}} F_i\right) = \sum_{i\in\mathbb{N}} m(F_i).$$

Mitta on funktio, joka riippuu sekä joukosta X että sigma-algebrasta $\mathcal{F}$. Joukon, sigma-algebran ja mitan kolmikkoa $(X, \mathcal{F}, m)$ voidaan kutsua mitta-avaruudeksi.

Määritelmä 2.1.5. *Todennäköisyysavaruus on mitta-avaruus* $(X, \mathcal{F}, m)$*, jolle pätee, että*

$$m(X) = 1.$$

Todennäköisyysavaruuden mittaa kutsutaan todennäköisyysmitaksi. Tätä mittaa merkitään usein kirjaimella $\mathbb{P}$. Todennäköisyysavaruudessa on myös tapana merkitä joukkoa X merkillä Ω. Määritellään vielä mitallinen funktio, jota käytetään satunnaismuuttujan määrittelyssä.

Määritelmä 2.1.6. *Oletetaan, että on olemassa mitalliset avaruudet* $(X, \mathcal{F})$ *ja* $(Y, \mathcal{G})$*. Funktio* $f : X \to Y$ *on mitallinen, jos jokaisen joukon* $B \in \mathcal{G}$ *alkukuva* $f^{-1}(B)$ *on mitallinen eli* $f^{-1}(B) \in \mathcal{F}$*.*

2.1.2 Satunnaismuuttuja

Seuraavaksi siirrytään mittateoriasta todennäköisyysteoriaan, joka aloitetaan määrittelemällä satunnaismuuttuja.

Määritelmä 2.1.7. *Satunnaismuuttuja on mitallinen funktio todennäköisyysavaruudesta* $(\Omega, \mathcal{F}, \mathbb{P})$ *mitalliseen avaruuteen.*

Määritelmä 2.1.8. *Reaaliarvoinen satunnaismuuttuja* X *on mitallinen funktio todennäköisyysavaruudesta* $(\Omega, \mathcal{F}, \mathbb{P})$ *reaaliavaruuteen eli* $X : \Omega \to \mathbb{R}$*.*

Kaikki tässä työssä käsiteltävät satunnaismuuttujat ovat reaaliarvoisia. Tästä eteenpäin tutkimuksessa satunnaismuuttujalla tarkoitetaan juuri reaaliarvoista satunnaismuuttujaa.

Määritelmä 2.1.9. *Satunnaismuuttujan* X *jakaumafunktio on* $F_X(x) = \mathbb{P}(X \leq x)$.

Esimerkki 2.1.10. Tasaisesti välillä $[0, 1]$ jakautuneen satunnaismuuttujan jakaumafunktio on

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1. \end{cases}$$

Satunnaismuuttujille voidaan myös määritellä lisää ominaisuuksia jakaumafunktion avulla. Ensimmäinen esiteltävä suure on satunnaismuuttujan odotusarvo.

Määritelmä 2.1.11. *Satunnaismuuttujan X odotusarvo $\mathbb{E}(X)$ on*

$$\mu_X = \mathbb{E}(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega),$$

mikäli integraali on määritelty.

Satunnaismuuttujan integraali tarkoittaa sitä, että integroidaan satunnaismuuttujan arvoa todennäköisyysmitan $\mathbb{P}$ suhteen. Eräs tapa ajatella odotusarvoa on se, että se kertoo satunnaismuuttujan arvojen keskiarvon. Kaikilla satunnaismuuttujilla ei kuitenkaan ole odotusarvoa, vaan määritelmän integraalia ei ole määritelty.

Yksi tärkeä ominaisuus on satunnaismuuttujan momenttiemäfunktio. Momenttiemäfunktion avulla voidaan laskea satunnaismuuttujan tunnuslukuja.

Määritelmä 2.1.12. *Satunnaismuuttujan X momenttiemäfunktio on*

$$M_X(s) = \mathbb{E}\left(e^{sX}\right),$$

silloin kun odotusarvo on olemassa.

Momenttiemäfunktio on yksi tapa esittää satunnaismuuttujan jakauma, sillä se on yksikäsitteinen jokaiselle satunnaismuuttujalle, jolla se on määritelty jollakin avoimella välillä. Momenttiemäfunktion avulla voidaan myös laskea satunnaismuuttujan momentit.

Määritelmä 2.1.13. *Satunnaismuuttujan X n-(origo)momentti on*

$$\mathbb{E}(X^n) = M_X^{(n)}(0).$$

Määritelmä 2.1.14. *Satunnaismuuttujan X n-keskusmomentti on*

$$\mu_n = \mathbb{E}\left[(X - \mathbb{E}(X))^n\right].$$

Seuraava esiteltävä suure on varianssi

Määritelmä 2.1.15. *Satunnaismuuttujan X varianssi, merkitään $\sigma_X^2 = \mathrm{Var}(X)$, on*

$$\sigma_X^2 = \mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Varianssi kertoo siitä, kuinka paljon satunnaismuuttujan arvot vaihtelevat eli kuinka paljon ne poikkeavat odotusarvosta. Mitä suurempi varianssi sitä suuremmalla todennäköisyydellä satunnaismuuttujan arvot painottuvat laajemmalle välille.

Satunnaismuuttujaa voidaan kuvata myös sen vinoudella, joka kertoo kummalle puolelle odotusarvoa satunnaismuuttujan arvot sijoittuvat. Vinouden laskemisessa käytetään kolmatta keskusmomenttia.

Määritelmä 2.1.16. *Satunnaismuuttujan X vinous, γ_X on*

$$\gamma_X = \frac{\mu_3}{\sigma^3} = \mathbb{E}\left(\frac{(X-\mu_X)^3}{\sigma_X^3}\right).$$

Jos vinous on positiivinen, niin jakauma on oikealle vino. Mikäli vinous on negatiivinen, niin jakauma on vasemmalle vino. Symmetristen muuttujien vinous on nolla, sillä niiden arvot ovat jakautuneet tasaisesti odotusarvon kummallekin puolelle. Satunnaismuuttujasta voidaan laskea myös huipukkuus, jonka laskemiseen käytetään neljättä momenttia.

Määritelmä 2.1.17. *Satunnaismuuttujan X huipukkuus on*

$$\kappa_X = \frac{\mu_4}{\sigma_X^4}.$$

Huipukkuudella kuvataan kuinka paljon häntäisyyttä jakaumassa on eli kuinka suuri osa jakaumasta on kaukana odotusarvosta.

2.2 Tilastotieteen teoriaa

Seuraavaksi esitellään tilastotieteestä tunnettuja teorioita ja määritelmiä, joita käytetään tässä työssä. Tilastotieteessä pyritään hahmottamaan ja mallintamaan todellisen maailman tapahtumia erilaisten jakaumien avulla. Tässä työssä mallinnettavana asiana on sairaanhoitoalueen kustannukset.

2.2.1 Lineaarinen riippuvuus

Tutkimuksessa käsitellään paljon toisistaan riippuvia muuttujia, joten on hyvä esitellä mitä tarkoittaa riippuvuus, ja varsinkin lineaarinen riippuvuus. Riippuvuus tarkoittaa sitä, että yhden muuttujan havaittu arvo vaikuttaa toisen, vielä havaitsemattoman muuttujan, jakaumaan. Lineaarisessa riippuvuudessa muuttujien vaihtelu on lineaarista.

Esimerkki 2.2.1. Jos X ja Y ovat toisistaan lineaarisesti riippuvia muuttujia parametrilla r, niin silloin yhtälö $X = rY + a$ pätee, missä a on myös jokin tunnettu vakio.

Lineaarinen riippuvuus on yleinen muuttujien välinen riippuvuus, jota esiintyy paljon tutkittaessa maailmassa esiintyviä ilmiöitä.

Esimerkki 2.2.2. Otetaan esimerkiksi auto, ja tutkitaan matkan ja matka-ajan välistä riippuvuutta, kun nopeus on vakio. Jos s on matka, t on aika ja v on nopeus, tällöin $s = tv$.

2.2.2 Aikasarja

Tässä osiossa esitellään lyhyesti aikasarjan määritelmä. Aikasarjaa käytetään esimerkkinä myöhemmin, kun esitellään tilastollisen mallin käsitettä. Tutkimuksessa aikasarjoja käytetään sai-

raanhoitoalueen kokonaiskustannuksien mallinnuksessa. Aikasarjojen teoriaa on esitelty tarkemmin Saikkosen luentomonisteessa [20].

Määritelmä 2.2.3. *Aikasarja on kokoelma havaintoja tietyltä yhtenäiseltä ajanjaksolta.*

Aikasarjaa merkitään yleensä ottaen, ja myös tässä tutkimuksessa, merkinnällä y_t, jossa $t \in \mathbb{N}$. Jos aikasarja on rajoitetulta ajanjaksolta, niin silloin $t = 0, 1, \ldots, T$. Yleisiä aikasarjojen aikavälejä kahden tapahtuman välillä ovat päivä, viikko, kuukausi, vuosineljännes ja vuosi. Tässä tutkimuksessa käsiteltävät aikasarjat ovat kuukausittain mitattuja aikasarjoja.

Yksinkertaisessa tapauksessa ajanjakson havainto y_t on reaaliluku, mutta se voi olla myös vektori. Tässä tutkimuksessa käsitellään vektoriarvoisia aikasarjoja, joissa $y_t \in \mathbb{R}^n, n \geq 2$. Tämän työn aikasarjoissa vektorien suuruus on rajattu, sillä suurimmat käsiteltävät dimensiot ovat luokkaa $y_t \in \mathbb{R}^6$. Aikasarjojen teoria kuitenkin yleistyy myös n-ulotteisille aikasarjoille. Mitä isompi aikasarjan ulottuvuus on, sitä isompi aineisto vaaditaan sen tarkkaan estimointiin. Myöhemmin esiteltävä oletus siitä, että parametrien lukumäärä on selvästi pienempi kuin havaittu aineisto, ei välttämättä päde, jos mallissa on monta muuttujaa.

Aikasarjoille voidaan laskea myös autokorrelaatiofunktio, joka kuvaa mallin havaintojen keskinäisiä korrelaatiota. Autokorrelaatiofunktion avulla voidaan laskea eri viipymien korrelaatiot. Moniulotteisille aikasarjoille voidaan laskea myös ristikorrelaatiofunktio, joka kuvaa eri muuttujien välisiä korrelaatioita. Tutkittavista aikasarjoista oletetaan yleensä, että ne ovat heikosti stationaarisia. Sen mukaan oletetaan, että aikasarjan odotusarvo ja kovarianssi ovat aikainvariantteja. Tällöin myös korrelaatiofunktiot ovat aikainvariantteja.

Määritelmä 2.2.4. *Oletetaan, että y_t on heikosti stationaarinen aikasarja, jolle pätee $y_t \in \mathbb{R}^n$, kaikilla $t \in \mathbb{N}$ ja jollakin $n \in \mathbb{N}, n \geq 2$. Määritellään kovarianssimatriisi viipymällä k seuraavasti*

$$\mathrm{Cov}(y_t, y_{t+k}) = \Gamma_k = \begin{bmatrix} \gamma_{11,k} & \cdots & \gamma_{1n,k} \\ \vdots & \ddots & \vdots \\ \gamma_{n1,k} & \cdots & \gamma_{nn,k} \end{bmatrix}.$$

Tällöin yksittäisen prosessin y_{at} autokorrelaatiofunktio on

$$\rho_{aa,k} = \frac{\gamma_{aa,k}}{\gamma_{aa,0}}, \quad \text{jossa } a = 1, \ldots, n.$$

Kahden eri prosessin y_{at} ja y_{bt} ristikorrelaatiofunktio on

$$\rho_{ab,k} = \frac{\gamma_{ab,k}}{\sqrt{\gamma_{aa,0}\gamma_{bb,0}}}, \quad \text{jossa } a, b = 1, \ldots, n, \text{ ja } a \neq b.$$

Auto- ja ristikorrelaatiot voidaan määritellä myös mallinnetun aikasarjan residuaaleille, jolloin niitä voidaan käyttää diagnostiikkatyökaluna.

2.2.3 Tilastollinen malli

Tilastollinen malli on matemaattinen konstruktio, jolla pyritään mallintamaan todennäköisyyksiä todellisten havaintojen takana. Tilastollisen mallin tavoitteena on usein pystyä ennustamaan tuntemattomia muuttujia tunnettujen havaintojen avulla. Mallia laskettaessa pyritään laskemaan todennäköisyysjakauma muuttujien takana. Kun tilastollinen malli rakennetaan, valitaan sopiva malli, jossa on parametri tai parametreja, joiden suuruus on tarkoitus laskea. Suuruutta ei kuitenkaan voida tietää tarkkaan, vaan niille lasketaan todennäköisimmät arvot. Parametrien todennäköisimmät arvot lasketaan käyttämällä tunnettua dataa, josta lasketaan millä parametreillä saatu aineisto on todennäköisin. Kun parametrien arvot ovat laskettu, niiden avulla voidaan ennustaa tulevia arvoja. Määritellään ensiksi tilastollinen malli tarkasti.

Määritelmä 2.2.5. *Tilastollinen malli on pari* $(\mathcal{X}, \mathcal{P})$*, missä* $\mathcal{X}$ *on joukko kaikista mahdollisista havainnoista, ja* $\mathcal{P}$ *on joukko todennäköisyysjakaumia, jotka ovat määritelty joukossa* $\mathcal{X}$*. Merkitään yksittäistä jakauma, jolla on parametrin arvo* θ*,* $P_\theta \in \mathcal{P}$*. Nyt voidaan kirjoittaa:* $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$*, missä* Θ *on joukko kaikista mahdollisista parametrin arvoista.*

Ajatellaan, että havaittu aineisto kuuluu joukkoon $\mathcal{X}$, ja että sillä on olemassa jokin todellinen jakauma. Tilastollisessa mallinuksessa tavoitteena on löytää joukosta $\mathcal{P}$ sellainen jakauma, joka on mahdollisimman lähellä todellista jakaumaa. Jakaumat ovat lähellä toisiaan, jos on suuri todennäköisyys saada jakaumista toisiaan vastaavia otoksia.

Yksi tässä tutkimuksessa esiintyvä tilastollinen malli on VAR(p)-malli. VAR(p)-malli on vektoriarvoisille aikasarjoille sovitettava malli. VAR(p)-mallin määrittelee yhtälö

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \text{iid}(0, \Omega) \quad (t \in \mathbb{Z}),$$

jossa oletetaan, että $y_t \in \mathbb{R}^n$ jollekin $n \in \mathbb{N}$. Mallissa $A_1, \ldots, A_p$ ovat parametrimatriiseja, jotka ovat kokoa $n \times n$ ja ε_t on $n \times 1$ matriisi joka kuvaa kunkin muuttujan virhetermiä.

2.2.4 Parametrien estimointi

Tilastollisessa mallinnuksessa on käytännössä mahdotonta löytää havaitun aineiston todellista jakaumaa. Tavoitteena on löytää sellainen parametrin arvo, jolla mallin jakauma olisi mahdollisimman lähellä todellista jakaumaa. Tämän takia puhutaan parametrien estimoinnista, sillä mallin tulokset ovat parhaimmillaankin vain approksimaatioita.

Parametrejä voidaan estimoida monilla eri tavoilla. Yksi yleisimmistä tavoista on pienimmän neliösumman menetelmä. Pienimmän neliösumman menetelmässä pyritään minimoimaan jäännöstermien neliöity summa. Jäännöstermit ovat mallin antamien ennusteiden ja todellisten arvojen erotus. Kun lasketaan niiden neliöiden summa, saadaan luku, joka kertoo kuinka paljon mallin antamat arvot poikkeavat todellisista arvoista.

Määritelmä 2.2.6. *Jäännösneliösumma on ei-negatiivinen reaaliluku, joka lasketaan seuraavalla kaavalla:*

$$JNS = \mathbb{E}((\hat{x} - x)^T(\hat{x} - x)),$$

missä $\hat{x}$ on mallin estimoidut arvot, ja x on mallin todelliset havaitut arvot. Termit $\hat{x}$ ja x ovat $N \times 1$ matriiseja, jossa N on tunnettujen arvojen määrä.

Esimerkki 2.2.7. Yksiulotteisessa tapauksessa, jolloin havaitut arvot x_n ja estimaatit $\hat{x}_n$ ovat reaalilukuja, voidaan kirjoittaa yksinkertaisesti $\hat{\varepsilon}_n = x_n - \hat{x}_n$, jolloin $\hat{\varepsilon}_n$ termejä kutsutaan mallin residuaaleiksi. Muuttuja n kuvaa havainnon järjestystä, ja sille pätee $n \in \{1, \cdots, N\}$. Silloin neliösumma voidaan kirjoittaa muodossa:

$$JNS = \sum_{n=1}^{N} \hat{\varepsilon}_n^2.$$

Pienimmän neliösumman käyttö on yksinkertainen ja tehokas tapa estimoida mallin parametrejä, ja se esiintyy useissa eri malleissa. Varsinkin lineaaristen regressiomallien estimoinnit pohjautuvat usein pienimmän neliösumman minimoimiseen.

2.2.5 Uskottavuusfunktio

Uskottavuusfunktiolla voidaan laskea tuntemattomien parametrien uskottavuus havaitun aineiston suhteen. Uskottavuusfunktio kuvaa kuinka uskottava valittu jakauma on suhteessa havaitun aineiston jakaumaan.

Määritelmä 2.2.8. *Uskottavuusfunktio on funktio $L(\theta)$, joka on määritelty parametrien arvoilla. Uskottavuusfunktion maksimikohta on suurimman uskottavuuden estimaatti.*

Mikäli uskottavuusfunktio tunnetaan ja se on jatkuvasti derivoituva, parhaimman parametrin valitseminen on helppoa. Silloin voidaan derivoida uskottavuusfunktiota ja ratkaista parametrin arvo derivaatan nollakohdassa. Tästä saadaan suurimman uskottavuuden estimaatti $\hat{\theta}$.

Esimerkki 2.2.9. Uskottavuusfunktio voidaan määritellä myös riippumaan havaitusta aineiston arvosta x. Tällöin uskottavuusfunktio on $L(\theta|x)$. Ja jos tutkitaan satunnaismuuttujaa X, jolla on tiheysfunktio f, niin uskottavuusfunktio voidaan määritellä myös $L(\theta|x) = f_\theta(x) = f(x|\theta)$.

2.2.6 Lineaarinen malli

Yksinkertaisessa lineaarisessa mallissa on selitettävä muuttuja Y ja yksi selittävä muuttuja x. Voidaan olettaa, että on olemassa useampi havainto selitettävästä muuttujasta. Tällöin havaintoaineistoa voidaan merkitä $Y_1, ..., Y_n$. Lisäksi oletetaan, että tiedetään jokaista havaintoa vastaava selittävän muuttujan arvo $x_1, ..., x_n$. Tavoitteena on se, että jos tiedämme selittävän

muuttujan x arvon, niin voimme ennustaa muuttujan Y arvon. Yksinkertaisessa tapauksessa tutkimme yhtälöä

$$Y_i = \alpha + \alpha_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

missä ϵ_i on ajanhetkestä i riippuva satunnainen virhetermi, jota ei voida havaita suoraan.

Normaaliapproksimoidussa lineaarisessa mallissa oletetaan, että virhetermit ovat itsenäisiä ja symmetrisesti normaalisti jakautuneita. Tällöin voidaan käyttää yleistä normaalijakauman merkintää: $N(\mu, \sigma^2)$, jossa μ on normaalijakauman odotusarvo ja σ^2 on normaalijakauman varianssi. Virhetermin oletus voidaan esittää nyt muodossa $\epsilon_i \sim N(\mu, \sigma^2)$.

Esimerkki 2.2.10. Oletetaan, että on olemassa havainnot kahdesta eri autokaupasta, jotka myyvät samaa autoa eri hinnalla. Oletetaan lisäksi, että tiedetään kuinka monta autoa ja millä hinnalla kumpikin autokauppa on myynyt. Tällöin voidaan muodostaa yksinkertainen lineaarinen malli, jolla ennustetaan autojen myyntimäärä sen hinnan perusteella. Merkitään, että H_i on hinta ja M_i on myytyjen autojen määrä yrityksissä $i = 1, 2$.

$$H_1 = 15000 \quad M_1 = 99$$

$$H_2 = 20000 \quad M_2 = 82$$

Näistä muodostettu lineaarinen malli on $M_i = \alpha + \alpha H_i + \epsilon_i$. Yksinkertaisella päättelyllä voidaan havaita, että mallissa suuremmalla hinnalla autoja myydään vähemmän. Seuraavassa osiossa lasketaan estimaatti parametrille α.

Yleisessä lineaarisessa mallissa selittävien muuttujien määrää ei ole rajattu, vaan niitä voi olla useampiakin. Merkitään seuraavaksi, että laskettavien parametrien lukumäärä on p, ja olemassa olevien havaintojen lukumäärä on n. Jos p on pienempi kuin havaintojen lukumäärä, niin mallin parametreja ei voida yleensä ottaen laskea yksikäsitteisesti. Ja mikäli p on lähes samansuuruinen kuin n, niin parametrien estimointi on epätarkkaa tai mahdotonta. Tämän takia yleensä oletetaan, että $p \ll n$, eli havaintoja oletetaan olevan paljon enemmän kuin parametrejä. Yleisen lineaarisen mallin yhtälö voidaan esittää muodossa:

$$Y_i = \beta + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i. \tag{2.1}$$

Yleisessä muodossa esitetyn lineaarisen mallin käsitteleminen voi olla haastavaa, mikäli on käytössä useita havaintoja, sillä silloin yhtälöryhmästä tulee iso. Tällöin yhtälöryhmän voi esittää parametrimuodossa, jossa

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \quad \text{ja} \quad \mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

Tällöin yleinen lineaariyhtälö voidaan kirjoittaa yksinkertaisessa matriisimuodossa

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Yleinen lineaarinen malli antaa mahdollisuudet monien erilaisten ilmiöiden tutkimiseen. Mallissa voidaan ottaa mukaan haluttaessa monta selittävää muuttujaa, mikä mahdollistaa monimutkaistenkin ilmiöiden selittämisen.

Esimerkki 2.2.11. Palataan aiempaan autoesimerkkiin, johon lisätään toinen selittävä muuttuja. Oletetaan, että tunnetaan aiemmin esiteltyjen suureiden lisäksi myös kaupan alueen asukkaiden keskimääräinen kuukausitulo, jota merkitään kirjaimella T. Laajennetaan aineistoa neljään havaintoon, jotka ovat:

$$H_1 = 15000 \quad M_1 = 99 \quad T_1 = 3000$$

$$H_2 = 20000 \quad M_2 = 82 \quad T_2 = 3000$$

$$H_3 = 15000 \quad M_3 = 115 \quad T_3 = 3500$$

$$H_4 = 20000 \quad M_4 = 99 \quad T_4 = 3500.$$

Nyt voidaan muodostaa ja ratkaista yhtälöryhmä, käyttämällä yhtälöä

$$M = \beta_1 H + \beta_2 T + \beta_0 + \epsilon.$$

Voidaan muodostaa neljän havainnon perusteella seuraavat yhtälöt:

$$\begin{cases} M_1 = \beta_1 H_1 + \beta_2 T_1 + \beta_0 + \epsilon_1 \\ M_2 = \beta_1 H_2 + \beta_2 T_2 + \beta_0 + \epsilon_2 \\ M_3 = \beta_1 H_3 + \beta_2 T_3 + \beta_0 + \epsilon_3 \\ M_4 = \beta_1 H_4 + \beta_2 T_4 + \beta_0 + \epsilon_4. \end{cases}$$

Käyttämällä pienimmän neliösumman menetelmää, jossa minimoidaan residuaalit ϵ_i, saadaan parametrien estimaateiksi:

$$\hat{\beta}_0 = 49.25 \quad \hat{\beta}_1 = -0.0033 \quad \hat{\beta}_2 = 0.033.$$

Kertoimista nähdään, että auton hinnan kerroin, eli β_1, on negatiivinen. Se tarkoittaa sitä, että auton hinnan kasvu vaikuttaa negatiivisesti auton myyntiin. Toisaalta asukkaiden keskitulon kerroin β_2 on positiivinen, joten keskitulon kasvu lisää myytyjen autojen määrää. Näillä kertoimilla voitaisiin pyrkiä ennustamaan mille tahansa auton hinnalle ja keskimääräiselle kuukausitulolle myytyjen autojen määrä.

2.2.7 Tilastollisia testejä

On olemassa paljon erilaisia tunnettuja tilastollisia testejä, joilla voidaan testata, onko jokin tilastollinen malli sopiva tietylle aineistolle. Testeillä voidaan tutkia, täyttääkö aineisto tietyt kriteerit tai onko aineistoilla jokin tietty ominaisuus. Tilastollisten testien käyttäminen on hyödyllistä, sillä niiden ominaisuudet ovat hyvin tunnettuja.

Esimerkki 2.2.12. Kolmogorov-Smirnov testissä tutkitaan yksiulotteisen aineiston empiiristä jakaumafunktiota $F_n(x)$. Jos halutaan tutkia onko aineisto jakautunut samoin kuin jokin tunnettu jakauma $F(x)$, niin voidaan laskea testisuure $T_{K-S} = \sup_x |F_n(x) - F(x)|$. Tämän testisuureen käyttö on perusteltu artikkelissa [10].

Testisuureet ovat usein jakautuneet jonkin tunnetun jakauman mukaisesti, jonka arvoihin suuretta verrataan. Yksi esimerkki on χ^2-jakauma.

Esimerkki 2.2.13. Jakauma χ^2 määritellään seuraavasti. Oletetaan, että $Z_1, \ldots, Z_k$ ovat riippumattomia standardinormaalijakautuneita satunnaismuuttujia. Näiden neliöiden summa on χ^2-jakautunut vapausasteella k eli

$$\sum_{i=1}^{k} Z_i^2 \sim \chi_k^2.$$

Jakauman χ^2 vapausaste valitaan aina tilanteen mukaan riippuen tutkittavien muuttujien tai parametrien määrästä. Jakaumaa käytetään esimerkiksi Portmanteau-testissä, joka esitellään seuraavaksi.

Esimerkki 2.2.14. Portmanteau-testillä voidaan tutkia ovatko moniulotteisen aikasarjan jäännöstermit riippumattomia, kuten mallioletukset vaativat. Testin nollahypoteesi on

$$H_0 : \mathbb{E}(\varepsilon_t \varepsilon_{t-i}^T) = 0, \quad i = 1, \ldots, h.$$

Eli nollahypoteesi on se, että residuaalit asteeseen h asti eivät ole auto- tai ristikorreloituneita. Eräs versio Portmanteau testisuureesta on

$$Q_h = T^2 \sum_{i=1}^{h} \frac{1}{T-i} \mathrm{tr}\left(S_i' S_0^{-1} S_i S_0^{-1}\right), \quad S_i = \sum_{t=1}^{T-i} \varepsilon_t \varepsilon_{t+i}'.$$

Yllä ε_t tarkoittaa mallin residuaalia, ja tr(A) operaattori tarkoittaa, että summataan matriisin lävistäjäalkiot yhteen. Jos VAR(p)-mallin virheet ovat korreloimattomia, niin testisuureelle pätee $Q_h \sim \chi^2$ vapausasteella $n^2(h-p)$. Jakauma χ^2 on yleisesti tunnettu jakauma, jonka arvot eri vapausasteilla tunnetaan. Näiden testien teoriaa esitellään tarkemmin Lütkepohlin kirjan [13] luvussa 4.

VAR(p)-mallin oletuksiin kuuluu myös se, että residuaalit ovat normaalijakautuneita. Tätä voidaan testa Lütkepohlin testillä.

Esimerkki 2.2.15. Lütkepohlin testissä lasketaan ensiksi aikasarjamallin residuaaleista standardisoidut versiot, joiden avulla lasketaan seuraavat testisuureet:

$$s_3^2 = \frac{T}{6}\left[\tfrac{1}{T}\textstyle\sum_{i=1}^{T}(\hat{\varepsilon}_{1i}^s)^3 \quad \cdots \quad \tfrac{1}{T}\sum_{i=1}^{T}(\hat{\varepsilon}_{ni}^s)^3\right] \cdot \begin{bmatrix} \frac{1}{T}\sum_{i=1}^{T}(\hat{\varepsilon}_{1i}^s)^3 \\ \vdots \\ \frac{1}{T}\sum_{i=1}^{T}(\hat{\varepsilon}_{ni}^s)^3 \end{bmatrix} \quad \text{ja}$$

$$s_4^2 = \frac{T}{24}\left(\begin{bmatrix} \frac{1}{T}\sum_{i=1}^{T}(\hat{\varepsilon}_{1i}^s)^4 \\ \vdots \\ \frac{1}{T}\sum_{i=1}^{T}(\hat{\varepsilon}_{ni}^s)^4 \end{bmatrix} - \begin{bmatrix} 3 \\ \vdots \\ 3 \end{bmatrix}\right)' \cdot \left(\begin{bmatrix} \frac{1}{T}\sum_{i=1}^{T}(\hat{\varepsilon}_{1i}^s)^4 \\ \vdots \\ \frac{1}{T}\sum_{i=1}^{T}(\hat{\varepsilon}_{ni}^s)^4 \end{bmatrix} - \begin{bmatrix} 3 \\ \vdots \\ 3 \end{bmatrix}\right).$$

Lisäksi lasketaan yhteistestisuure:

$$JB_K^L = s_3^2 + s_4^2.$$

Yhtälöissä $\hat{\varepsilon}_{ji}^s$ ovat estimoituja standardisoituja residuaaleja ja T on aineiston koko. Testien teoria ja standardisoitujen residuaalien kaava on esitelty Lütkepohlin kirjassa [13] luvussa 4. Testisuureista s_3^2 mittaa residuaalien vinoutta, s_4^2 mittaa residuaalien huipukkuutta ja JB_K^L mittaa molempia. Näiden testisuureiden arvoja verrataan χ^2 jakaumiin, sopivilla vapausasteilla, siten että suureita s_3^2 ja s_4^2 verrataan jakaumaan $\chi^2(n)$ ja suuretta JB_L^K jakaumaan $\chi^2(2n)$. Mikäli testisuureen arvo on liian kaukana χ^2-jakauman arvosta, hylätään nollahypoteesi residuaalien normaaliudesta.

2.2.8 Yliselittäminen

Yliselittämisellä tarkoitetaan tilastollisessa mallintamisessa sitä, että jos malliin otetaan mahdollisimman paljon muuttujia mukaan, voidaan saada puhtaasti selitysasteeltaan hyvä malli, mutta mallilla ennustaminen ei välttämättä ole tarkkaa. Tämä johtuu siitä, että jollekin muuttujalle saatetaan laskea parametrin arvo satunnaisvaihtelun takia, vaikka oikeasti muuttuja ei vaikutta ollenkaan mallin arvoihin. Eli toisin sanoen osa mallin parametreista on tarpeettomia mallinnuksen kannalta.

Schwarzin informaatiokriteeri tai uudemmalta nimeltään Bayesian information criterion, lyhyemmin BIC, on menetelmä, jolla voidaan rajoittaa mallia useamman muuttujan käyttämisestä. Tarkoituksena on se, että jos muuttuja tuo malliin vähemmän selitysastetta kuin mitä kriteerille ominainen sakkofunktio sakottaa uuden muuttujan lisäämisestä, niin muuttujaa ei valita malliin. Eri yhteyksissä BIC:llä on erilaisia määritelmiä, mutta tässä tutkimuksessa sitä käytetään valitsemaan lineaarisen regressiomallin muuttujien lukumäärä ja VAR(p)-mallien valintakriteerinä. BIC arvon perusteella päätellään mikä VAR(p)-mallin aste p olisi paras.

Kun muodostetaan VAR(p)-mallia, joudutaan valitsemaan malliin sopivat muuttujat ja mallin aste. Mallin aste valitaan minimoimalla funktio

$$BIC(p) = \log\left(\det(\tilde{\Omega}_\varepsilon(p))\right) + \frac{\log(T)}{T}pn^2.$$

Yhtälössä merkintä det tarkoittaa matriisin determinanttia, n on mallin muuttujien lukumäärä, T on aineiston koko ja p on testattava mallin aste. Yhtälössä $\tilde{\Omega}_\varepsilon(p)$ on estimoitu kovarianssimatriisi, joka estimoidaan kaavalla

$$\tilde{\Omega}_\varepsilon(p) = \frac{1}{T}\sum_{i=1}^{T} \hat{\varepsilon}_i \hat{\varepsilon}_i'.$$

Kaavassa esiintyvä residuaali estimaatti $\hat{\varepsilon}_i$ riippuu mallin asteesta p, sillä residuaalin kaava on

$$\hat{\varepsilon}_i = y_i - \hat{A}_1 y_{i-1} - \ldots - \hat{A}_p y_{i-p}.$$

Funktion $BIC(n)$ minimikohdasta saadaan BIC:n mukaan paras estimaatti mallin asteelle p. Funktion käytön perustelu on esitelty tarkemmin Schwarzin artikkelissa [21]. Varsinainen BIC:n ominainen osuus funktiossa on $\log T$ jälkimmäisen termin osoittajassa. Vertailun vuoksi esitellään myös Akaiken informaatiokriteerin funktio, joka on esitelty tarkemmin Akaiken artikkelissa [1],

$$AIC(p) = \log\left(\det(\tilde{\Omega}_\varepsilon(p))\right) + \frac{2}{T}pn^2.$$

Kolmantena kriteerifunktiona käytetään Hannah-Quinn -informaatiokriteeriä, jonka funktio on:

$$HQ(p) = \log\left(\det(\tilde{\Omega}_\varepsilon(p))\right) + \frac{2\log(\log(T))}{T}pn^2.$$

Funktioissa muuttujien selitykset ovat samat kuin BIC-funktiossa.

Eri mallinvalintakriteerit saattavat antaa eri mallin asteen vastaukseksi, sillä niille ominaiset sakkofunktiot eroavat toisistaan. Mallien tulkinta ja mallin asteen valitseminen on aina tutkijan vastuulla, joka viime kädessä tekee päätöksen mallin asteen valinnasta. Tässä tutkimuksessa pyritään rajoittamaan mallin kokoa, sillä käytettävissä olevan aineiston määrä rajoittaa paramterien estimointia.

3 Aineisto

Käytössä on HUSin tarjoama aineisto vuosilta 2011-2019, jossa on kattavasti tallennettuna tietoja HUSin toiminnasta. Aineistossa on eritelty kuukausitasolla HUSin erilaiset kulutyypit kuten henkilöstökustannus, lääkekustannus ja esimerkiksi rakennuksiin liittyvät kustannukset. Yksittäisten potilaiden laskujen jakauman määrittelemisessä aineistona on käytetty anonyymeja tietoja yksittäisten potilaiden hoitojen laskutuksesta. Lisäksi apuna käytetään myös Terveyden ja Hyvinvoinnin laitoksen tuottamia aineistoja väestömääristä.

Aineiston tutkimiseen on käytetty pääasiassa R-ohjelmointikieltä [17]. Suurin osa tarvittavista funktioista ja kuvaajista on koodattu tätä tutkimusta varten. Valmiit paketit on mainittu tulosten yhteydessä, mikäli niitä on käytetty. Kuvaajien laatimiseen on käytetty apuna Cairopakettia [23]. Aikasarjamallinnuksessa ja -estimoinnissa on käytetty JMulti-ohjelmaa [14].

3.1 Aineiston tarkka kuvaus

Tutkimuksessa on käytössä HUSin tarjoamaa aineistoa toimintakuluista, diagnooseista, henkilöstöistä ja arkipäivien lukumäärästä. Tärkeimpänä aineistona on tilikartta, jossa on eritelty jokaiselta kuukaudelta eri kustannus- ja tuottolajit. Näistä tutkitaan tarkimmin kohtaa T4 Toimintakulut, jossa on eriteltynä ne kustannuslajit, joista tässä tutkimuksessa ollaan kiinnostuneita. Taulukossa on lisäksi listattuna T3 Toimintatuotot, T6 Rahoitustuotot ja -kulut sekä T7 Poistot ja arvonalentumiset. Toimintatuotot riippuvat vahvasti kustannuksista, joten niiden tutkiminen jätetään pois. Rahoitustuotot taas ovat säännöllisesti erittäin pieniä, joten niiden vaikutus on käytännössä olematon. Poistot ja arvonalentumiset taas ovat hyvin säännönmukaisia ja yleensä ottaen hyvin tiedossa, joten nekin jätetään pois tarkastelusta.

Tilikartan lisäksi aineistona käytetään diagnoositaulukkoa, jossa on kuukausittain eritelty eri diagnoosien lukumäärä ja niihin liittyviä suureita: tuotelukumäärät, hoitopäivien lukumäärät, DRG-pisteet ja kustannukset. Diagnoosit ovat nimetty ICD-10 luokituksen kolmemerkkisten koodien mukaan, jotka on yhdistelty HUSin aineistossa erilaisiin diagnoosiryhmiin. Keskimäärin kuukaudessa on diagnooseja hieman alle 200 diagnoosiryhmästä. Tätä tarkemmalla diagnoositasolla tutkiminen ei olisi järkevää, sillä silloin yksittäisen diagnoosiryhmän tuotelukumäärä olisi jo liian pieni, jotta sen tutkiminen olisi tilastollisesti mielekästä.

Tutkimuksessa tutkitaan myös aineistoa henkilöstöstä, jossa on kuukausittain eroteltuna henkilöstön lukumäärä ja heidän tekemänsä henkilötyövuodet. Koska aineistoa tutkitaan kuukausitasolla, käytämme myös tietoa arkipäivien lukumäärästä kuukausittain. Tämä johtuu siitä, että elektiivistä hoitoa eli ei-kiireellistä hoitoa, jonka hoitopäivä voidaan valita vapaasti, pyritään toteuttamaan lähinnä arkipäivisin.

Lisäksi on vielä tuotteet-nimellä kutsuttu taulukko, jossa on kuukausittain eritelty eri tuotelajeittain kustannukset, DRG-pisteet, hoitopäivät, potilaslukumäärät, tuotelukumäärät, hoidossa olleiden henkilöiden lukumäärä ja hoitojaksojen lukumäärä. Hoitopäivät ja hoitojaksot

liittyvät ainoastaan vuodehoitoihin. Vuodehoidoissa ei ole poliklinikkakäyntejä ollenkaan eli niitä on ainoastaan muissa tuotelajeissa.

DRG-pisteet ovat sairaanhoitopiirin tapa kuvata kunkin toimenpiteen vaativuutta tai resurssien tarvetta. Jokaiselle sairaanhoitopiirissä tuotettavalle hoidolle, joita kutsutaan tuotteiksi, määritellään vuosittain DRG-pistearvo. DRG-pistearvot määritellään yleensä ottaen edellisen vuoden toteutuneiden kustannusten pohjalta. Nämä pisteet ovat yksi selkeimmistä tavoista tutkia todellisesti toteutettua hoitoa, sillä eri potilaiden vaativuustaso voi vaihdella suuresti.

Käytössä on myös tiedot Hyvinkään sairaanhoitopiirin kuntien, Hyvinkään, Järvenpään, Mäntsälän, Nurmijärven ja Tuusulan, väestötiedoista, jotka on haettu Sotkanetistä [12]. Palvelusta on haettu sekä väkiluvut vuoden viimeisenä päivänä vuosilta 2012-2018, että ennusteet vuodelle 2025. Lisäksi palvelusta on myös haettu yli 65-vuotiaiden ihmisten osuus väestöstä samoina ajankohtina.

Tutkimuksen viimeisessä osiossa tarkasteltaviin yksittäisten potilaiden kustannuksiin käytetään taulukkoa, johon on koottu vuosilta 2015-2019 yksittäisten potilaiden kustannukset vuositasolla. Aineistossa on yhteensä noin 150 tuhatta potilasta.

Tämän aineiston avulla pyritään löytämään merkittävimpiä ja oleellisimpia kustannustekijöihin vaikuttavia tekijöitä, ja pyrkiä muodostamaan malli kustannusten ennustamiseen. Lisäksi pyritään mallintamaan yksittäisen potilaan kustannusten jakaumaa.

3.2 Aineiston tunnuslukujen esittely

Aineiston esittely aloitetaan visuaalisesti tutkimalla kuvaajia eri kustannuslajeista. Tutkimuksessa käytetään kuukausista juoksevaa numerointia, jossa vuoden 2012 tammikuu on numero 1, ja vuoden 2019 marraskuu on numero 95. Ensimmäisenä tutkitaan yleistä kuvaajaa eri kustannuslajeista, joka on esitelty kuvassa 3.1.

Kuvasta 3.1 huomataan, että suurin osa kustannuslajeista pysyy tarkasteltavalla ajanjaksolla likimain samalla tasolla. Kuvaajasta nähdään myös se, että kokonaiskustannukset kasvavat pitkällä aikavälillä, ja kustannusten määrä vaihtelee kuukausittain paljon. Kokonaiskustannuksista voidaan myös havaita säännöllistä kausivaihtelua: siinä on lähes jokaisena joulukuuna ollut pieni notkahdus alaspäin, jonka jälkeen alkuvuodesta kustannukset nousevat taas ylöspäin.

Muutokset ovat kuitenkin hieman epäsäännöllisiä, ja vaihtelua on paljon eri kuukausina, joten siirrytään tutkimaan eri kustannuslajeja. Kuvan 3.1 kuvaaja osoittaa selvästi, että kahdella suurimmalla yksittäisellä kustannuslajilla on aivan erilaiset kuvaajat. Niiden kausivaihtelut poikkeavat selvästi toisistaan. Kokonaiskustannuksissa tämä vaihtelu peittyy, sillä huiput ovat kustannuslajista riippuen erilaiset varsinkin joulukuussa.

Kustannuslajeista suurimmat ovat selvästi T41 palkat ja palkkiot, T430 sairaanhoidollisten palvelujen osto, T454 lääkkeet ja hoitotarvikkeet, T42 henkilö-sivukulut ja T434 toimisto- ja asiantuntijapalvelut. Sama asia voidaan todeta myös laskemalla kuukausittaiset keskiarvot jokaisesta kustannuslajista. Kymmenen suurimman kustannuslajin keskiarvot ovat laskettuina

Kuva 3.1: Kuvaaja kaikista kustannuslajeista

taulukkoon 3.1.

Taulukosta 3.1 nähdään nyt, että suhteellisesti isoimmat erot kustannuslajien suuruuksissa ovat ensimmäiseksi ja toiseksi suurimman välillä sekä seitsemänneksi ja kahdeksanneksi suurimman välillä. Palkkojen ja palkkioiden suuruus tulee myös selvästi esille, se on selvästi isoin yksittäinen kustannuslaji.

3.3 Merkittävimmät kustannuslajit

Seuraavaksi tutustutaan kuvaajaan, joka on esitelty kuvassa 3.2. Siihen on piirrettynä ainoastaan seitsemän suurinta kustannuslajia. Kuvan 3.2 kuvaajasta nähdään kuvaa 3.1 selvemmin, että eri kustannuslajeissa on erilaiset kausivaihtelut. Suurimmassa kustannuslajissa, palkoissa ja palkkioissa on selvästi piikit ylöspäin joulukuiden kohdalla. Sairaanhoidollisten palvelujen ostoissa taas piikit ovat alaspäin joulukuiden kohdalla. Mielenkiintoista on se, että suurin osa

Taulukko 3.1: Taulukko kustannuslajien kuukausittaisista keskiarvoista.

T41 PALKAT JA PALKKIOT	3 792 587
T430 SAIRAANHOIDOLLISTEN PALVELUJEN OSTOT	1 693 485
T454 LÄÄKKEET JA HOITOTARVIKKEET	1 052 418
T42 HENKILÖ–SIVUKULUT	874 816
T434 TOIMISTO- JA ASIANTUNTIJAPALVELUT	717 852
T480 VUOKRAT	563 795
T438 PUHTAANAPITO- JA PESULAPALVELUT	437 342
T447 MUUT PALVELUT	129 935
T441 MAJOITUS- JA RAVITSEMISPALVELUT	129 421
T442 MATKUSTUS- JA KULJETUSPALVELUT	75 340

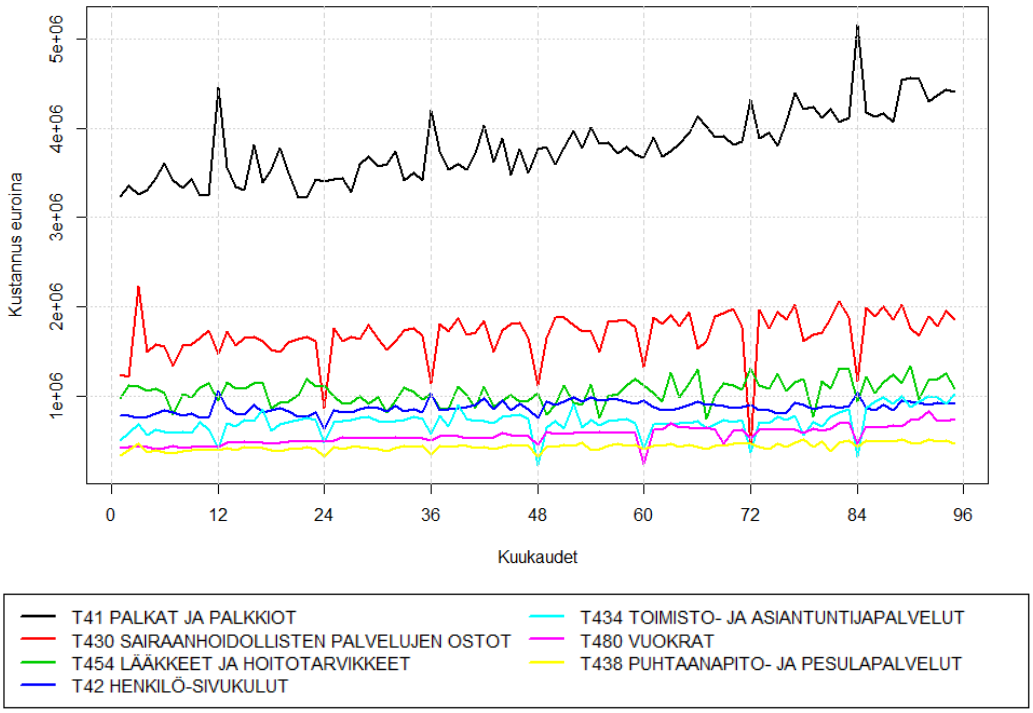
kustannusten noususta näyttää johtuvan palkkojen ja palkkioiden noususta, sillä ne ovat tarkasteluajanjaksolla nousseet noin miljoona euroa. Lääkkeiden ja hoitotarvikkeiden kustannukset eivät ole kasvaneet kovinkaan paljon, niissä esiintyy kuitenkin paljon vaihtelua kuukausittain. Muiden kustannuslajien suuruus on kasvanut vuodesta 2013 vuoteen 2019, mutta kasvu ei ole yhtä selvästi havaittavissa, sillä kustannuslajit ovat itsessään pienempiä.

Vuokrien sekä puhtaanapito- ja pesulapalveluiden kustannukset näyttävät noudattavan hyvin lineaarista kaavaa, pois lukien pienet notkahdukset joulukuisin. Täten näiden ennustaminen tulee olemaan yksinkertaista, sillä niiden voidaan olettaa kasvavan lineaarisesti. Muilla kustannuslajeilla on enemmän vaihtelua, joten niiden osalta tarvitaan tarkempaa analyysiä.

3.4 Tuotetyyppitaulukon esittely

Tuotetyyppitaulukkoon on taulukoitu kustannukset tuotetyypeittäin kuukausitasolla. Taulukosta on mahdollista tutkia eri hoitotyyppien kustannuksia. Taulukon tuotetyypeistä merkittävimmät ja suurimmat ovat vuodehoito, päivystys, ensikäynti, uusintakäynti, sarjahoitokäynti, hoitokäynti, hoitopuhelu, päiväkirurgia, terveyskeskus päivystyskäynti ja hoitokirje. Näistä on vielä valikoitu viisi oleellisinta kuvan 3.3 kuvaajaan tarkastelemalla kunkin tuotetyypin keskiarvoja. Vuodehoito esitetään erikseen kuvan 3.5 kuvaajassa, sillä se on huomattavasti suurempi kuin muut tuotelajit. Mielenkiintoista kuvan 3.3 kuvaajassa on se, että päivystyksellisten tuotteiden kustannukset ovat kasvaneet selvästi, vaikka muut kustannukset ovat pysyneet lähes samalla tasolla kun verrataan vuosien 2013 ja 2019 tilanteita. Tämä voisi olla yksi selittävä tekijä kustannusten kasvulle. Tutkimme sitä tarkemmin vertaamalla päivystyksen kustannusten osuutta kokonaiskustannuksesta sekä päivystyksen DRG-pisteiden osuutta kaikista DRG-pisteistä, jotka näkyvät kuvassa 3.4.

Tuotetyyppien kustannuksista mielenkiintoista on myös se, että muiden kuin päivystyksel-

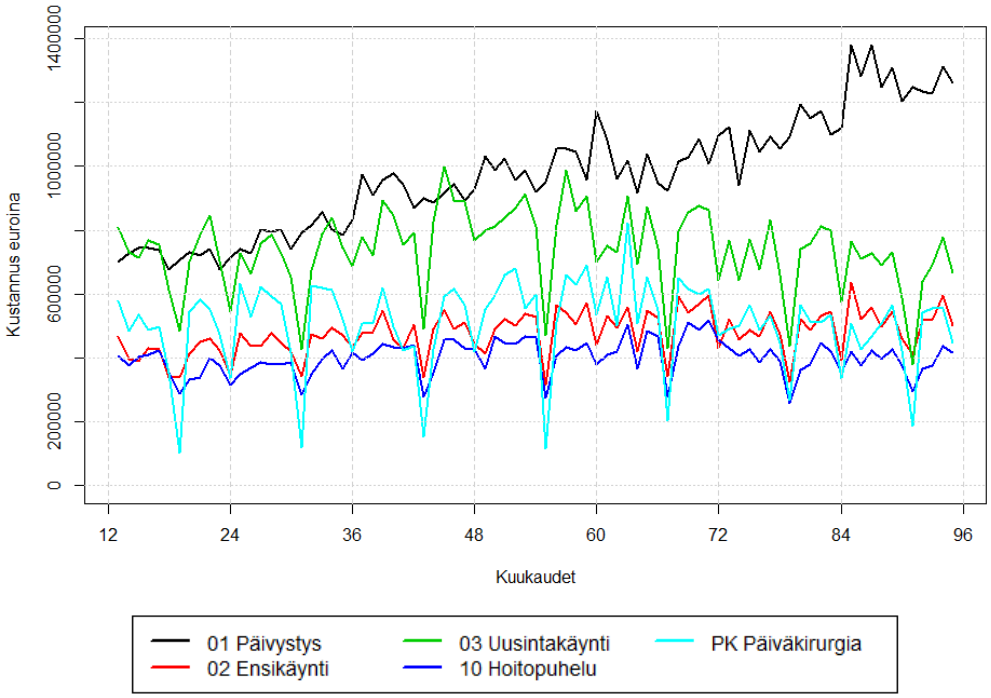


Kuva 3.2: Kuvaaja suurimmista kustannuslajeista. Kausittaiset vaihtelut näkyvät selvästi erisuuntaisina piikkeinä.

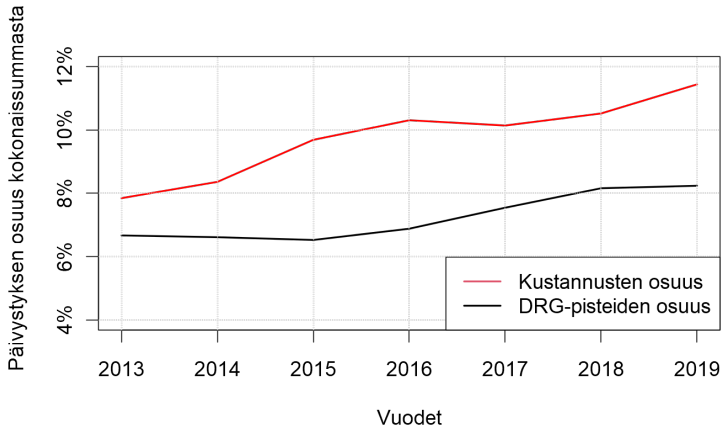
listen hoitojen kustannukset eivät ole kasvaneet, vaikka kuudessa vuodessa inflaation vaikutus saattaisi näkyä kustannuksissa. Lisäksi kokonaiskustannukset ovat kasvaneet tasaisesti, mutta esimerkiksi uusintakäynnin ja päiväkirurgian kustannukset ovat pysyneet samalla tasolla tai jopa laskeneet kuukauden 60 eli vuoden 2016 joulukuun jälkeen.

Päivystyksellisten tuotteiden kustannuksia tutkiessa voidaan summata vuoden aikana päivystyksessä tuotetut DRG-pisteet ja verrata niiden osuutta koko vuoden aikana tuotetuista DRG-pisteistä. Tämä osuus on kuvan 3.4 mukaan kasvanut tasaisesti vuosien saatossa, ja vaikka kasvu on ollut prosenttiyksiköissä pientä, se on suhteellisesti merkittävä muutos. Kustannuksia tutkittaessa vastaavalla tavalla muutos on ollut vielä suurempaa. Tämä on merkittävä muutos, sillä päivystyksessä tuotetun DRG-pisteen hinta on ollut joka vuosi huomattavasti kalliimpi kuin yleinen DRG-pisteen hinnan keskiarvo.

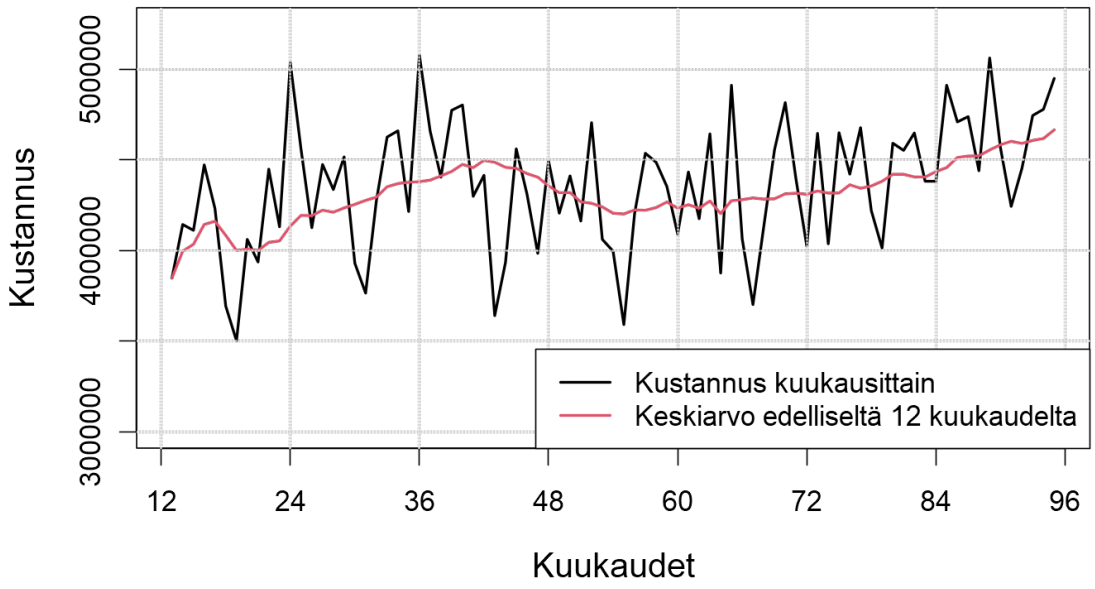
Vuodehoitojen kustannukset ovat piirrettynä kuvaajaan 3.5. Vuodehoitojen kustannukset vaihtelevat todella paljon kuukausittain, ja sen takia kuvaajaan on piirretty liukuva keskiar-



Kuva 3.3: Kuukausittaiset kustannukset eroteltuina tuotetyyppien mukaan. Päivystyksen kustannukset kasvavat selvästi tarkastellulla ajanjaksolla.



Kuva 3.4: Kuvaaja päivystyksen osuudesta kokonaiskustannuksista. Osuus kasvaa selvästi vuosien mittaan.



Kuva 3.5: Vuodehoidon kustannukset kuukausittain. Musta viiva kuvaa vuodehoidon kustannuksia kuukausitasolla, ja niiden liukuva kahdentoista kuukauden keskiarvo on merkattu punaisella viivalla.

vo, jotta kustannusten muutosta on helpompi seurata. Liukuva keskiarvo on laskettu siten, että lasketaan keskiarvo enintään kahdestatoista edellisestä kuukaudesta. Kustannukset ovat selvästi suurempia kuin muut tuotelajit, mikä johtuu vuodehoidon vaatimista suurista resursseista. Vuodehoidon kustannusten kuukausittainen keskiarvo oli yli 4 miljoonaa euroa, kun taas päivystyksellisten tuotteiden kuukausittainen keskiarvo oli hieman alle miljoona euroa. Mielenkiintoinen havainto on vuoden 2015 ja 2016 välillä tapahtunut lasku kustannusten keskiarvossa, jonka jälkeen kustannukset ovat taas nousseet.

4 Sairaanhoitoalueen kokonaiskustannuksen estimointi

4.1 Tutkimusmenetelmän esittely

Kokonaiskustannusten muutoksiin merkittävimpien tekijöiden selvittämiseen käytetään lineaarista mallia. Ensiksi aineistosta tulee valita sopivimmat muuttujat lineaariseen malliin. Valinta tehdään käyttämällä hyväksi tietämystä sairaanhoitopiirin toiminnasta ja tutkimalla muuttujien välisiä korrelaatioita sekä niiden vaikutusta kokonaiskustannuksiin. Lopuksi ennustetaan aikasarjamallilla kokonaiskustannuksia käyttämällä hyödyksi aiemmassa vaiheessa tärkeiksi havaittuja muuttujia.

4.2 Tutkimusmenetelmän perustelu

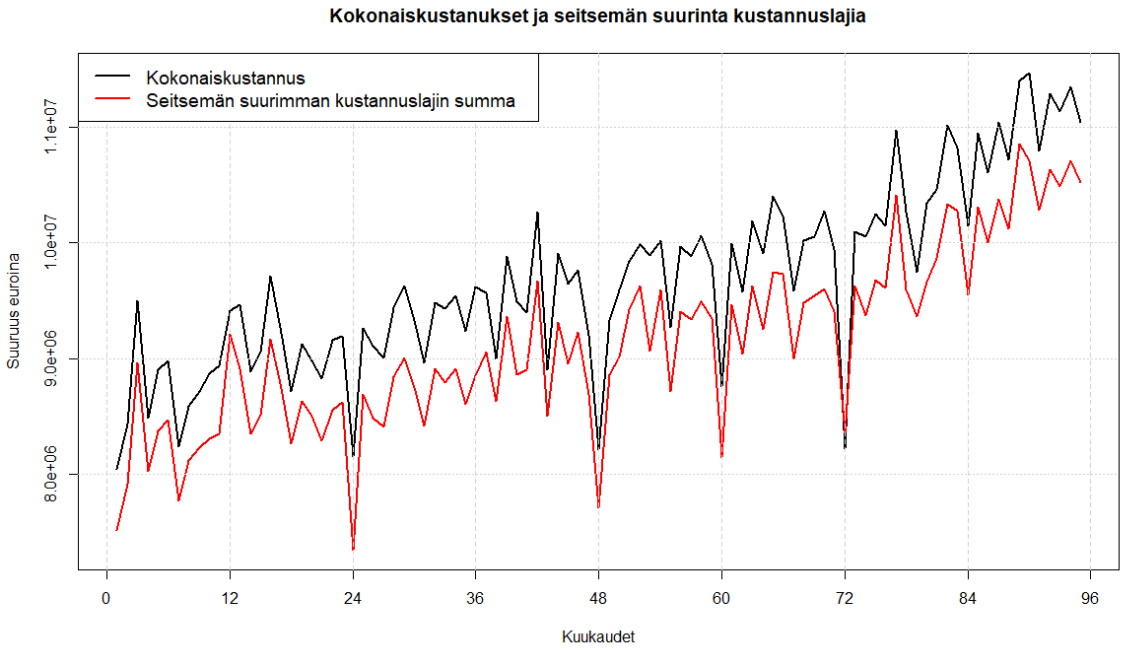
Perustellaan lyhyesti, miksi lineaarista mallia voidaan käyttää tutkimuksessa, ja mitkä ovat mallin oletukset.

Lineaarisen mallin yksi oletuksista on se, että virhetermit noudattavat normaalijakaumaa. Tutkimuksessa normaaliutta tärkempää on jakauman symmetrisyys, sillä lineaarisen mallin teoria pätee suurimmilta osin myös ilman oletusta normaaliudesta. Perustelu jakaumien symmetrisyydestä voidaan tehdä tutkimalla mallien residuaaleja, joiden todetaan olevan tarpeeksi hyviä mallin käytettävyyden kannalta. Virheiden oletetaan myös olevan riippumattomia, minkä voidaan olettaa seuraavan siitä, että vaikka potilaiden määrillä ja kustannuksilla on riippuvuuksia kuukausien välillä, niin kokonaiskustannusten satunnaistermien voidaan olettaa olevan ainakin lähes riippumattomia. Tätä myös testataan aikasarjamallinnuksen diagnostiikassa.

Toinen oletus on se, että kokonaiskustannus on lineaarinen yhdistelmä selittävistä muuttujista, minkä voidaan olettaa olevan tarpeeksi lähellä totuutta. Selittävien muuttujien oletetaan lisäksi olevan tarkkoja arvoja, mikä historiaa tutkittaessa on realistinen oletus, joten oletuksen voidaan sanoa pätevän. Lisäksi oletetaan, että mallin parametrimatriisilla on yksikäsitteinen ratkaisu.

4.3 Muuttujien valinta

Rajoitamme alustavan tarkastelun seitsemään suurimpaan kustannuslajiin, koska on yksinkertaisempaa käsitellä pientä määrää kustannuslajeja. Voidaan myös osoittaa, että pienimmät kustannuslajit eivät ole merkityksellisiä kokonaiskustannusten kannalta. Perustelemme tämän rajoitetun tarkastelun riittävyyden tilastollisten menetelmien avulla. Mallin rajaamisella saamme myös laskettua mallin dimensiota, mikä tarkentaa parametrien estimointia. Alustavasti voidaan todeta, että keskimäärin seitsemän suurinta kustannuslajia muodostavat 94% kustannuksista. Kuvassa 4.1 on esitelty sekä kokonaiskustannusten että seitsemän suurimman kustannuslajin summa kuukausittain. Liitteiden kuvaan A.1 taas on piirretty näiden erotus. Kuvasta 4.1



Kuva 4.1: Kokonaiskustannus ja seitsemän suurimman kustannuslajin summa.

nähdään, että kummatkin viivat liikkuvat lähes vastaavalla tavalla. Kuvasta A.1 huomataan myös, että kokonaiskustannusten ja seitsemän suurimman kustannuslajin ero on suhteellisen tasainen.

Myös tilastollisella menetelmällä voidaan osoittaa, että seitsemän isoimman kustannuslajin summa on jakautunut lähes vastaavasti kuin kokonaiskustannukset. Koska nämä kaksi suuretta eivät ole riippumattomia, esimerkissä 2.2.12 esiteltyä Kolmogorov-Smirnov -testiä ei voi käyttää. Sen sijaan tutkitaan niiden erotuksen ja siihen sovitetun kiinteän tason residuaalivektoria. Halutaan osoittaa, että residuaalit ovat autokorreloimattomia ja normaalisti jakautuneita. Tällöin voidaan sanoa, että seitsemän suurinta kustannuslajia ovat samoin jakautuneita kuin kokonaiskustannukset. Tällöin voidaan asettaa kiinteä luku, joka kuvaa kaikkien muiden kustannuslajien suuruutta.

Residuaalien autokorrelaatiofunktion arvot, jotka on esitelty kuvassa A.2, eivät ylitä luottamustasoa. Lisäksi kvantiilikuvaajassa, joka on kuvassa A.3, suurin osa arvoista on luottamusvälin sisällä. Kuvaajan perusteella jakauman hännät ovat kuitenkin hieman normaalijakaumaa paksumpia eli todennäköisyysjakauma ei täysin vastaa normaalijakaumaa. Jakauma vaikuttaa kuitenkin symmetriseltä, mikä on oleellisempaa kuin jakauman normaalius estimoinnin robustisuuden eli häiriöherkkyyden kannalta. Voisi olla mahdollista etsiä jokin jakauma, joka sopii paremmin residuaaleihin, mutta tässä tutkimuksessa sivuutetaan tämä.

Täten voidaan siis perustellusti valita kokonaiskustannusten ennustamiseen vain seitsemän suurinta kustannuslajia, mikä yksinkertaistaa ennustemallin rakentamista ja syy-seuraussuhteiden löytämistä. Pienemmissä kustannuslajeissa on myös enemmän satunnaisvaihtelua, joka saattaisi aiheuttaa epätarkkuutta ennusteeseen. Lisäksi ennustaminen helpottuu, kun erotetaan eri kustannuslajit, sillä niiden kuukausittaiset vaihtelut poikkeavat toisistaan huomattavasti.

Aloitamme analyysin yhdistämällä sopivat kustannuslajit. Aluksi summataan palkat ja palkkiot sekä henkilö-sivukulut yhteen. Henkilö-sivukulut liittyvät olennaisesti palkkojen ja palkkioiden suuruuteen, koska molemmat riippuvat henkilöstön lukumäärästä ja heidän tekemästään työmäärästä. Henkilö-sivukulujen tulisi pitkällä aikavälillä olla tietty osuus henkilöstön kokonaispalkasta, ja vaikka tämä ei kuukausitasolla päde, niin ennustamisen kannalta on loogista summata nämä yhteen. Lisäksi kuvaajasta nähdään, että kustannuslajien vaihtelut ovat hyvin vastaavanlaisia, molemmissa on piikit vastaavissa kohdissa varsinkin kuukausina 12, 36 ja 84 eli joulukuut 2012, 2014 ja 2018. Kustannuslajit vaihtelevat tarpeeksi samalla tavalla, jotta niiden summaaminen on perusteltua.

Toiset kaksi kustannuslajia, jotka voidaan summata yhteen, ovat sairaanhoidollisten palvelujen ja toimisto- ja asiantuntijapalveluiden ostot. Nämä kaksi kustannuslajia linkittyvät selvästi toisiinsa, kuten voidaan huomata jokaisen joulukuun kohdalla, molemmissa kuvaajissa on selvä piikki alaspäin. Ja on perusteltua olettaa, että mikäli sairaanhoitoalue ostaa sairaanhoidollisia palveluja, niin sairaanhoitoalue ostaa myös mahdollisesti näihin sairaanhoidollisiin palveluihin liittyviä toimisto- ja asiantuntijapalveluita.

Toinen perustelu näiden kustannuslajien yhdistämiselle voidaan löytää tutkimalla kustannuslajien otoskorrelaatioita. Kun tarkastellaan taulukon 4.1 kustannuslajien otoskorrelaatiomatriisia, niin nähdään kustannuslajien keskinäiset otoskorrelaatiot.

Ensimmäisenä taulukosta 4.1 huomataan, että palkat ja palkkiot korreloivat eniten henkilöstö-sivukulujen kanssa, mikä on vastaava yhdistelmä kuin aikaisemmin. Ja vaikka palkat ja palkkiot korreloivat hieman vuokrien kanssa, niin henkilöstö-sivukulut näyttävät korreloivan ainoastaan palkkojen ja palkkioiden kanssa. Lisäksi sairaanhoidollisten palvelujen osto näyttää korreloivan toimisto- ja asiantuntijapalveluiden kanssa, joten tämä yhdistelmä on myös perustellusti valittu. Mielenkiintoista otoskorrelaatiomatriisissa on myös se, että lääkkeiden ja hoitotarvikkeiden kustannukset eivät näytä korreloivan vahvasti minkään muun kustannuslajin kanssa. Myös mallissa olevat likimain lineaarisesti kasvavat kustannuslajit, vuokrat ja pesulakustannukset, näyttävät korreloivan keskenään.

Täten tutkittavien muuttujien määrää voidaan vähentää kustannuslajeja yhdistämällä. Jatkossa palkkojen, palkkioiden ja henkilöstö-sivukulujen summaa kutsutaan henkilöstökuluiksi. Sairaanhoito-, toimisto- ja asiantuntijapalveluiden ostoja voidaan kutsua lyhyemmin palvelujen ostoina. Tämä yhdistäminen tehdään sen takia, jotta erilaisten kustannuslajien määrä pysyisi rajattuna, ja mallissa olisi vähemmän komponentteja. Tällöin mallin muodostaminen on yksinkertaisempaa, ja kun yhdistäminen on tehty maltillisesti ja perustelluilla päätöksillä, niin tilastollisen analyysin luotettavuuden ei pitäisi muuttua merkittävästi. Jos yhdistämisen

Taulukko 4.1: Taulukko kustannuslajien otoskorrelaatioista ja kustannuslajien nimet.

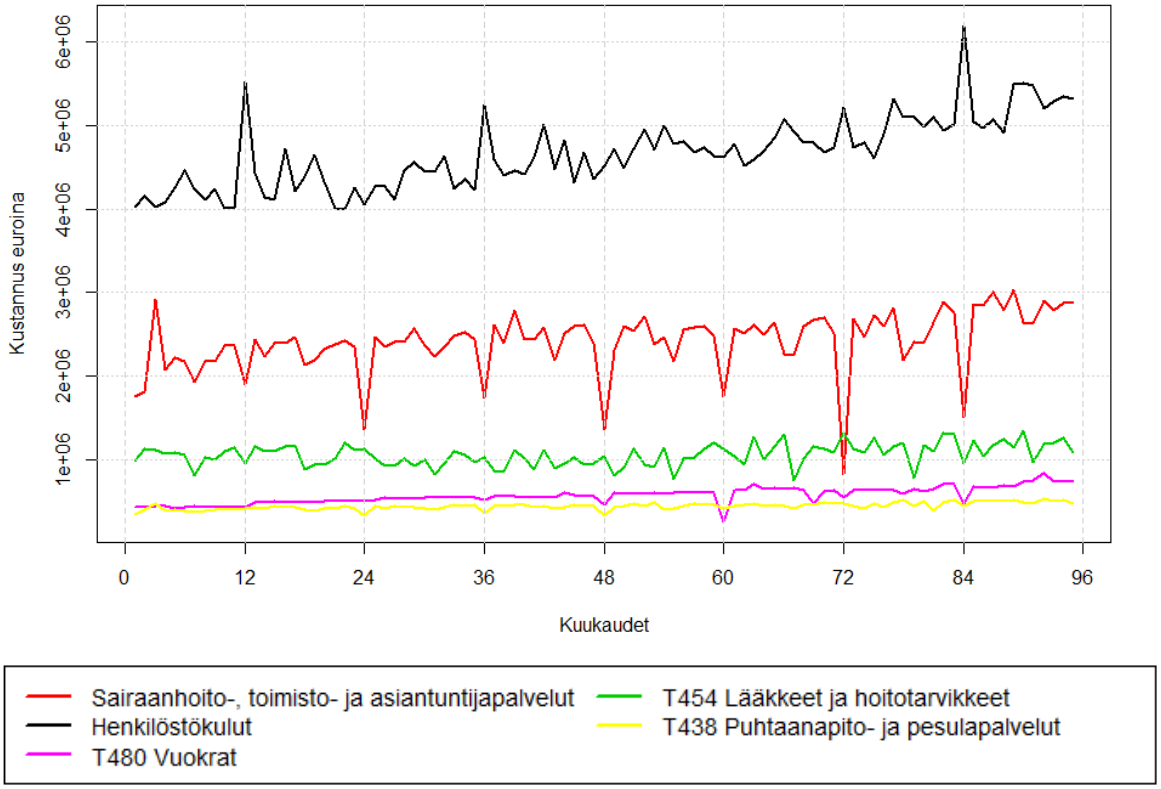
Laji	T41	T430	T454	T42	T434	T480	T438
T41	1	0.11	0.23	0.69	0.21	0.61	0.49
T430	0.11	1	0.17	0.13	0.69	0.50	0.60
T454	0.23	0.17	1	-0.04	0.21	0.24	0.43
T42	0.69	0.13	-0.04	1	0.12	0.32	0.34
T434	0.21	0.69	0.21	0.12	1	0.65	0.62
T480	0.61	0.50	0.24	0.32	0.65	1	0.69
T438	0.49	0.60	0.43	0.34	0.62	0.69	1

T41	PALKAT JA PALKKIOT
T430	SAIRAANHOIDOLLISTEN PALVELUJEN OSTOT
T454	LÄÄKKEET JA HOITOTARVIKKEET
T42	HENKILÖ–SIVUKULUT
T434	TOIMISTO- JA ASIANTUNTIJAPALVELUT
T480	VUOKRAT
T438	PUHTAANAPITO- JA PESULAPALVELUT
T447	MUUT PALVELUT
T441	MAJOITUS- JA RAVITSEMISPALVELUT
T442	MATKUSTUS- JA KULJETUSPALVELUT

vaikutusta halutaan tutkia tarkemmin tai eri näkökulmasta, voidaan kustannuslajit hajottaa takaisin seitsemään eri osaan, ja ennustaa niitä yksitellen.

Kuvan 4.2 kuvaajasta näkee, että tutkittavana on lineaarisesti kasvavien kustannuslajien lisäksi kolme erilaista kustannuslajia. Lääkkeiden ja hoitotarvikkeiden kustannusten vaihtelu näyttää erittäin satunnaiselta, mutta sitä voidaan ennustaa käyttämällä keskiarvoa, ja ennuste tarkentuu, kun tutkitaan ennustetta pidemmällä aikavälillä, sillä pieni kuukausittainen vaihtelu tasaantuu, kun tutkitaan useampaa kuukautta. Tutkimme myös mahdollisia muuttujia, jotka vaihtelevat yhtä paljon kuin tarvikkeiden kustannukset, ja saattaisivat selittää kuukausittaista vaihtelua.

Palvelujen ostoissa on näkyvillä selkeä vuosittainen trendi, jossa kustannukset tippuvat joulukuussa, nousevat alkuvuodesta, tippuvat vähän kesällä, nousevat loppusyksyyn, kunnes taas putoavat joulukuuhun. Vaikuttaisi siltä, että lomakaudella sairaalassa tuotetaan vähemmän palveluita, joten ostot jäävät pienemmiksi. Aikasarjamallinnuksessa tällaisen selkeän kausivaihtelun ennustaminen on mahdollista, sillä jokaiselle kuukaudelle pystytään määrittelemään vakio. Näin voidaan mallintaa jokaisen kuukauden ominaista vaihtelua. Piikkien tulevaisuuden kokojen ennustaminen on haasteellista, sillä saatavilla ei ole kuin seitsemän kokonaisen vuoden



Kuva 4.2: Kuvaaja yhdistetyistä kustannuslajeista.

tiedot, joten kunkin kuukauden vakion määrittelemiseen on käytössä vain seitsemän havaintoa. Palveluiden ostoissa näyttää myös olevan pientä nousua tarkastellulla aikavälillä.

Henkilöstökulujen vaihtelu on kaikista kustannuslajeista epäselvintä, joten sen ennustaminen on todennäköisesti vaikeinta. Joissakin joulukuissa on henkilöstökuluissa selvä piikki ylöspäin, kuten kuukausina 12, 36 ja 84, jotka vastaavat vuosia 2012, 2014 ja 2018. Muina vuosina piikkejä ei ole, tai piikit ovat selvästi pienempiä kuin muina vuosina. Tämä tarkoittaa sitä, että ennustaessa pitää olla tarkkana, ettei yliarvioi joulukuun vaikutusta, mutta se täytyy kuitenkin huomioida. Toinen yleinen vaihtelutrendi henkilöstökuluissa on se, että kesäkuukausina on piikki ylöspäin. Tämä kesäkuukausien piikki voidaan ottaa huomioon, kun rakennetaan mallia ennusteelle.

4.4 Lineaariset mallit

Tässä osiossa rakennamme lineaarisen mallin kullekin viidestä kustannuslajista. Aloitamme lineaarisen mallin rakentamisen helpoimmista muuttujista eli vuokrista ja puhtaanapitopalveluista. Näille molemmille voidaan laskea yksinkertaiset kertoimet ajan suhteen. Tällöin kustannukset voidaan ennustaa taulukon 4.2 mukaan. Taulukossa 4.2 kuukausikerroin tarkoittaa kuukauden juoksevaa numeroa, joka alkaa vuoden 2012 tammikuusta.

Seuraavaksi siirrymme vaikeammin ennustettaviin kustannuslajeihin. Käytämme mallin ennustamisessa R-koodia, jolla voidaan laskea eri muuttujien osajoukoille ennusteen tarkkuus, ja estimoida sitten Schwarzin informaatiokriteerin avulla mitkä muuttujat olisivat parhaita. Ensiksi tarvitaan taulukko kaikista muuttujista, joita halutaan käyttää ja joista karsitaan sopivimmat muuttujat varsinaiseen malliin.

Valitsemme mahdollisten muuttujien taulukkoon seuraavat muuttujat: kuukauden numero, DRG-pisteiden summa, väestön lukumäärä, arkipäivien lukumäärä, hoitopäivien lukumäärä, kaikki tehdyt henkilötyövuodet yhteensä, poliklinikkakäynnit, syöpähoitojen lukumäärä, nivelsairaushoitojen lukumäärä ja kasvain diagnoosiin liittyvien hoitojen lukumäärä. Käyttämällä tätä taulukkoa laskemme nyt lineaarisen mallin sairaanhoidollisten palvelujen ostoille.

Kuvan A.4 kuvaajassa on esitelty sekä BIC:n, että yleisen selitysasteen arvot eri muuttujien lukumäärillä. BIC:n minimikohdassa on paras ennuste lineaariselle mallille. Vaaka-akselilla on muuttujien lukumäärä, ja pystyakselilla on mallin testiarvo. BIC:tä käytettäessä arvo pyritään minimoimaan, mutta selitysasteen tapauksessa isompi arvo on parempi. Huomataan, että selitysasteen arvo nousee jatkuvasti, kun taas BIC:n tapauksessa kuvaajassa on vaihtelua eri muuttujien lukumäärillä. Tämä juontuu aiemmin esitellystä yliselittämisongelmasta. Valitsemme nyt BIC:n mukaisen seitsemän muuttujan mallin.

Taulukkoon 4.3 on koottu lääkekustannusten mallin kertoimet. Kuukausikerroin on saman juoksevan kuukausinumeron kerroin kuin mikä on esitelty aiemmin. Mallista huomataan, että kustannukset nousevat ajan myötä, vaikka väestön lukumäärän kerroin on negatiivinen. Muiden tekijöiden vaikutus kustannuksiin on selkeä, kukin muuttuja nostaa kustannuksia tietyn arvon verran.

Seuraavaksi voidaan laskea vastaavat ennusteet myös henkilöstökuluille ja sairaanhoidollisten palvelujen ostoille. Käytämme näiden ennustamiseen pohjana samaa mahdollisten muuttujien taulukkoa kuin lääkkeiden ennustamiseen, mutta lisäämme siihen henkilöstön tekemät työvuodet eriteltyinä erikseen muille työntekijöille, hoitajille ja lääkäreille. Valitsemme BIC:tä käyttäen sopivimmat muuttujat malliin. Tulokset on esitelty taulukoissa 4.4 ja 4.5.

Taulukoista 4.4 ja 4.5 nähdään, että yleisesti ottaen väestön lukumäärä vaikuttaa negatiivisesti kustannuksiin, vaikka kuukauden järjestysnumero vaikuttaa positiivisesti. Tämä kertoo siitä, että kustannukset kasvavat ajan myötä. Väestön lukumäärän kerroin voisi implikoida sitä, että mitä enemmän väestöä on, sen tehokkaampaa sairaanhoidon tuotanto on. Toisaalta kuukausi termin kasvu on lineaarista, kun taas väestön kasvu on lähes lineaarista, joten nämä

Taulukko 4.2: Taulukko vuokrien ja pesulakustannuksien malleista.

Pesulakustannukset	Vuokrakustannukset	Yhteensä
387000 + 1060·Kuukausi	438000 + 2690·Kuukausi	824000 + 3740·Kuukausi

Taulukko 4.3: Taulukko lääkekustannusten mallista.

Lääkekustannukset	
Vakiotermi	50300000
Kuukausi	27200
DRG-pisteiden summa	35.3
Väestön lukumäärä	-275
Arkipäivät	23100
Nivelsairaus tuotteiden lukumäärä	447
Kesäkuu	168000
Joulukuu	173000

kaksi muuttujaa ovat lähes samoja. Tällöin mallin kertoimet eivät välttämättä kerro totuutta yksittäisen muuttujan vaikutuksesta.

Huomionarvoista on myös se, että taulukoiden 4.4 ja 4.5 malleissa joulukuu on valittu merkittäväksi muuttujaksi, vaikkakin se on henkilöstökuluissa positiivinen lisäys, ja palvelujen ostoissa negatiivinen. Joulukuu on siis selvästi poikkeava kuukausi eri kustannuslajien kannalta, mikä varmasti johtuu osaltaan joulunpyhistä ja ihmisten käyttäytymisestä niiden aikana. Henkilöstökulujen kertoimiin kuuluu henkilöstöstä kertovista tekijöistä ainoastaan muun henkilökunnan tekemät henkilötyövuodet, mikä voi vaikuttaa hieman erikoiselta. Yleinen logiikka voisi sanoa, että lääkäreiden ja hoitajien lukumäärä vaikuttaisi myös henkilöstökustannuksiin. Todennäköinen selitys on se, että muun henkilökunnan tekemät henkilötyövuodet seuraavat paremmin henkilöstökustannusten vaihtelua kuin muiden henkilöstölajien tekemät henkilötyövuodet.

4.5 Aikasarjamallinnus

Aikasarjamallinnuksessa halutaan konstruoida hyvä VAR(p)-malli kokonaiskustannusten ennustamiseen. Aineiston analysoinnissa käytetään apuna JMulti-ohjelmaa, jolla voidaan analysoida aikasarjoja ja estimoida malleja. VAR(p)-mallin valintaprosessissa tulee valita sopivat muuttujat, sekä laskea näille sopiva mallin aste p. Kun valinta on suoritettu, tulee vielä varmistaa, että malli täyttää lähtöoletukset, ja että estimoidut parametrit käyttäytyvät hyvin.

Taulukko 4.4: Taulukko henkilöstökustannusten mallista.

Henkilöstökulut	
Vakiotermi	86800000
Väestön lukumäärä	-457
Kuukausi	57000
Muun hlk htv	137000
Joulukuu	354000

Taulukko 4.5: Taulukko palvelujen ostojen kustannusten mallista.

Palvelujen ostot	
Vakiotermi	76500000
Väestön lukumäärä	-413
Kuukausi	45200
Hoitopäivät	170
Poliklinikkakäynnit	48.1
Joulukuu	-1090000

Aloitetaan mallin rakennus karsimalla muuttujia. Käytössä on rajattu aineisto, joten mallissa ei voi olla liikaa parametreja, jotta estimointi olisi tarkkaa. Aluksi rajataan mahdolliset muuttujat muuttujiin, jotka ovat mukana jossain aiemmin estimoidussa lineaarisessa mallissa. Käytettävänä on silti liikaa muuttujia, jotta estimointi olisi mielekästä. Karsitaan pois sellaiset muuttujat, jotka ovat herkempiä satunnaisuudelle, tai niiden vaikutuksen arvioidaan olevan vähäinen. Lisäksi käytetään apuna mallinvalintakriteeriä, jolla voidaan laskea numeerinen arvo kunkin mallin hyvyydelle, ja näitä vertailemalla voidaan päättää mitä mallia käytämme. Näissä kriteereissä pienempi arvo on parempi.

Ensimmäisenä karsitaan pois muun henkilökunnan tekemät henkilötyövuodet, sillä sen arvo on pieni ja siten altis heilahteluille. Lisäksi sen vaikutus kokonaiskustannuksiin vaikuttaa epäselvältä. Toinen pieniä arvoja saava muuttuja on nivelsairaus tuotteiden lukumäärä, joka jätetään myös pois. Seuraavaksi voidaan nyt laskea valintakriteerien arvot mallille, jossa on muuttujina kokonaiskustannus, hoitopäivät, poliklinikkakäynnit, väestön lukumäärä ja DRG-pisteiden summa. Käytetään osiossa 2.2.7 esiteltyjä kriteerifunktioita laskemaan mallien sopivuutta. Tulokset eri kriteerifunktioiden minimeistä on koottu taulukkoon 4.6.

Sama lasku voidaan tehdä myös mallille, josta on jätetty DRG-pisteet pois. DRG-pisteiden ongelmallinen luonne on se, että jokaisen sairaalan tuottaman tuotteen DRG-piste arvot lasketaan uudelleen vuosittain, eikä kahden eri vuoden välillä tuotetut DRG-pisteet ole välttämättä vertailukelpoisia. Tämän takia tutkitaan mallia, josta DRG-pisteet on jätetty pois. Tämän mallin kriteerifunktion arvot on esitelty taulukossa 4.7.

Koska tavoitteena on minimoida kriteerifunktion arvot, voidaan taulukoiden 4.7 ja 4.6 tuloksista huomata, että on perusteltua valita malliksi neljän muuttujan malli. Malli on hyvä myös siitä näkökulmasta, että hoitopäivien ja poliklinikkakäyntien lukumäärä ovat mielenkiintoisia suureita, joiden ennustaminen on mielekästä.

Mallin asteeksi valitaan 2, joka saadaan laskemalla BIC-kriteerifunktion arvot eri asteille. Seuraavaksi tulee testata, että malli on toimiva, ja että residuaalit täyttävät mallin vaatimat oletukset. Käytämme jo esiteltyä Portmanteau-testiä tarkastaaksemme, että mallissa ei ole

Taulukko 4.6: Viiden muuttujan mallin valintakriteerin arvot.

AIC:	64.6
BIC:	68.1
HQ:	66.0

Taulukko 4.7: Neljän muuttujan mallin kriteerifunktioiden arvot.

AIC:	51.8
BIC:	54.3
HQ:	52.8

Taulukko 4.8: Testi residuaalien normaaliudesta.

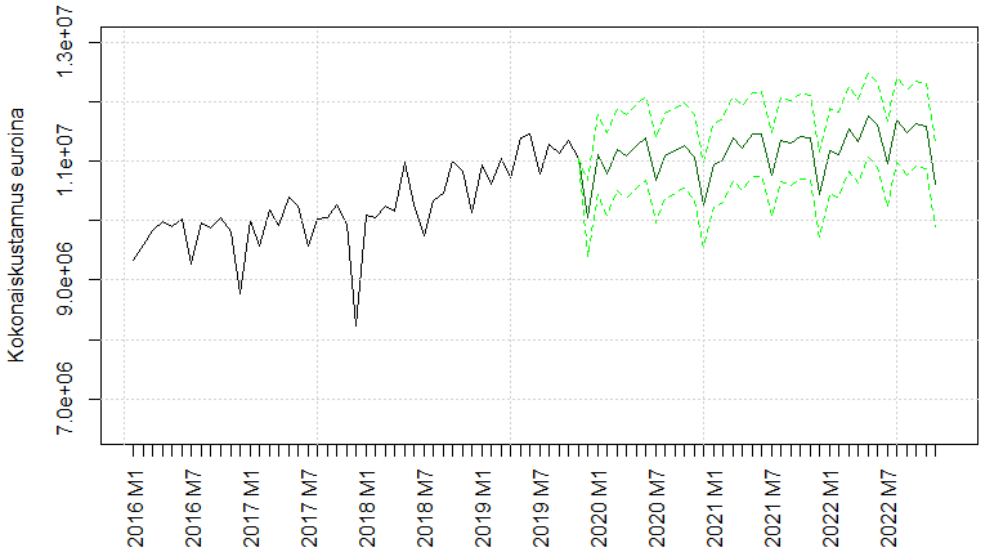
Yhteistestisuure:	382
p-arvo:	0.000
Vinoustestisuure:	3.62
p-arvo:	0.458
Huipukkuustestisuure:	378
p-arvo:	0.000

näyttöä lähtöoletusta mallin residuaalien korreloimattomuutta vastaan. Laskemme testin arvon asteella 7, jolloin p-arvoksi tulee 0.136, mikä tarkoittaa sitä, että nollahypoteesia ei hylätä. Mikäli p-arvo olisi ollut alle 0.05, nollahypoteesi olisi hylätty 95-prosentin luottamustasolla. Lisäksi voidaan tutkia residuaalien auto- ja ristikorrelaatiokuvaajia, jotka ovat esiteltyinä liitteiden kuvassa A.6.

Breusch-Godfrey LM -testi antaa myös tulokseksi asteella 7, että residuaalit eivät olisi autokorreloituneita. Voimme siis todeta, että ei ole näyttöä sitä oletusta vastaan, että residuaalit eivät ole auto- tai ristikorreloituneita. Täten siis mallin lähtöoletus virheiden riippumattomuudesta pätee.

Toinen mallin lähtöoletus on se, että virheet ovat normaalisti jakautuneita. Tätä voidaan testata esimerkiksi Jacque-Bera testillä. Testin tuloksena hoitopäivien ja poliklinikkakäyntien residuaalit ovat normaalisti jakautuneita. Vaikka huolestuttavasti kokonaiskustannusten residuaalit eivät ole tämän testin mukaan normaalisti jakautuneita, voidaan asiaa tutkia tarkemmin lisätesteillä.

Yksi vaihtoehto on käyttää esimerkissä 2.2.15 esiteltyä Lütkepohlin testiä. Testin tulokset ovat lueteltuina taulukossa 4.8. Taulukossa on ensimmäisena laskettu yhteistestisuure ja siihen liittyvä p-arvo. Seuraavana on vinoustestin arvo ja siihen liittyvä p-arvo, jonka jälkeen on huipukkuustestisuure ja sen p-arvo. Taulukosta 4.8 nähdään, että testi hylkää nollahypoteesin residuaalien normaaliudesta, sillä yhteistestin p-arvo on pieni. Testin arvo virheen vinoudelle jättää kuitenkin nollahypoteesin voimaan, sillä sen p-arvo on 0.458, mikä on suurempi kuin 0.05, jolla hypoteesi hylättäisiin 95-prosentin luottamustasolla. Täten voidaan siis päätellä, että residuaalien arvojen huiput ovat liian suuria olemaan normaalisti jakautuneita. Tällainen



Kuva 4.3: Kokonaiskustannusten ennuste vuosille 2020-2022. Ennuste on piirretty tummanvihreällä viivalla ja 95%-luottamusväli on piirretty vaaleanvihreällä katkoviivalla.

residuaalien huipukkuus ei ole niin ongelmallista kuin se, että residuaalien vinous poikkeaisi normaalijakaumasta. Mallin robustisuuden takia huipukkuus ei ole vakava ongelma, sillä estimaattoreiden jakaumat ovat asymptoottisesti samat kuin tilanteessa, missä residuaalit ovat normaalijakautuneita.

Residuaalien tiheysfunktiot voidaan myös mallintaa ja verrata niitä normaalijakauman tiheysfunktioon. Todetaan, että kuvaajat vaikuttavat olevan hyvin lähellä normaalijakaumaa, ainoastaan väestönkasvun jakauma näyttäisi olevan keskittynyt enemmän nollan ympärille kuin normaalijakaumassa. Tämä ei kuitenkaan ole merkittävä syy epäillä mallin luotettavuutta.

Mallin kertoimien stabiiliutta voidaan myös tutkia. Tällä tarkoitetaan sitä, että tutkitaan miten mallin kertoimet muuttuvat kun lisätään havaintoja joista kertoimet lasketaan. Mikäli kertoimessa tapahtuu suuria muutoksia viimeisten havaintojen kohdalla, se kertoisi siitä, että muuttujassa tapahtuu jonkin rakenteellinen muutos ja kertoisi mallin toimimattomuudesta. Näiden rekursiivisesti laskettujen kertoimien kuvaajat on esitelty liitteen kuvassa A.6. Mallin kohdalla ei näy mitään viitteitä siitä, että olisi syytä epäillä mallin epäsopivuutta.

Impulssivasteanalyysillä voidaan tutkia kunkin muuttujan vaikutusta toisiin muuttujiin eri vasteajoilla. Tämä ohitetaan tässä tutkimuksessa toteamalla vain, että väestön lukumäärän kasvu kasvattaa hoitopäivien ja poliklinikkakäyntien lukumäärää. Lisäksi todetaan, että impulssivasteissa ei ole mitään erikoista, mikä saattaisi implikoida mallin epäpätevyydestä.

Kun malli on todettu tarpeeksi päteväksi, voidaan sen avulla ennustaa tulevaisuuden arvoja. Asetetaan ennustushorisontti 37 kuukauteen, mikä tarkoittaa, että ennuste arvioi muuttujien arvot vuoden 2022 loppuun asti. Kuvassa 4.3 on esitelty tämän ennusteen tulokset kokonaiskustannuksille, ennusteen lisäksi kuvaajaan on piirretty myös 95-prosentin luottamusväli. Huomataan, että kokonaiskustannuksissa näyttää olevan hienoista kasvua tulevina vuosina. Varsinkin joulukuun kausivaihtelun alimmat laskut näyttäisivät nousevan korkeammalle. Kokonaisuudessaan kuvan 4.3 ennusteen mukaan kustannukset kasvavat vain vähän.

Poliklinikkakäyntien ja hoitopäivien lukumäärien ennusteet ovat esiteltyinä liitteissä, kuvissa A.7 ja A.8. Poliklinikkakäyntien lukumäärän ennustetaan kasvavan tulevaisuudessa, kun taas hoitopäivien lukumäärä näyttäisi hieman laskevan. Tietysti eri muuttujien ja mallin valinnoilla voitaisiin saada erilaiset tulokset, mutta tutkimamme malli vaikuttaa olevan hyvä diagnostiikan perusteella. Väestön lukumäärä kasvaa mallin ennusteessa lineaarisesti tulevaisuudessa. Väestörekisterikeskuksen ennustetta vuodelle 2025 käytettäessä väestön kasvu hidastuisi selvästi havainnoitujen kuukausien jälkeen, joten mallin ennustama väestön lukumäärä saattaisi olla lähempänä totuutta.

5 Yksittäisten potilaiden kustannusten analysointi

Tässä luvussa tutkitaan yksittäisten potilaiden kustannuksia, ja niistä erityisesti pienen todennäköisyyden tapauksia, joiden kustannukset ovat kuitenkin suuret. Pienellä todennäköisyydellä tapahtuvat tapahtumat voivat olla todella suuria, ja ne saattavat aiheuttaa merkittävän kustannuserän sairaanhoitopiirille. Tällaisten ennustaminen on tärkeää, jotta pystytään varautumaan yllättäviin kustannuksiin. Pienten todennäköisyyksien tapahtumista voidaan tutkia sekä niiden suuruutta että niiden tapahtumistodennäköisyyttä.

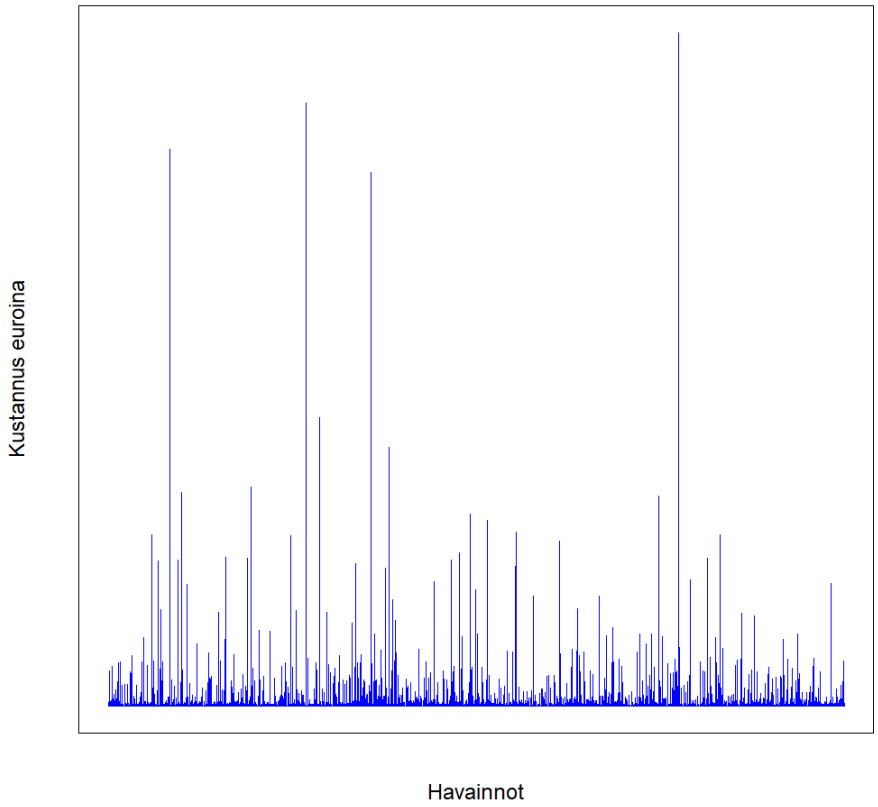
Aineistona käsitellään potilaiden kustannuksia vuosilta 2015-2019 Hyvinkään sairaanhoitoalueella. Kunkin potilaan kustannukset on summattu yhteen näiltä vuosilta. Näistä puhutaan jatkossa laskuina. Aineistosta olisi perimmäisenä tarkoituksena pyrkiä ennustamaan tulevien potilaiden laskut. Toinen tavoite on pyrkiä arvioimaan, millaiset kustannukset saattaisivat keskimäärin olla siinä tilanteessa, että pienen todennäköisyyden tapahtuma tapahtuu.

Sopivan mallin tai jakauman löytäminen on monivaiheinen prosessi, jossa aluksi rajataan tutkittava mallijoukko. Alle 10 eri jakauman visuaalinen tarkastelu verrattuna otokseen on vielä mahdollista, mutta jos tutkittavana olisi yli 100 erilaista jakaumaa, kuvaajien tulkinta olisi käytännössä mahdotonta. Kun sopiva tutkittava jakaumajoukko on löytynyt, näistä voidaan karsia pois huonoiten aineistoon sopivat jakaumat. Parhaiten aineistoon sopivien jakaumien toimivuutta voidaan tarkastella tarkemmin tekemällä tilastollisia testejä ja laskemalla erilaisia diagnostiikkasuureita, kuten keskineliövirhe. Lisäksi voidaan simuloida valituista kandidaattijakaumista otoksia, ja verrata niitä havaittuun aineistoon. Näistä vertailuista olisi tavoitteena löytää yksi jakauma, jota ei aineiston pohjalta voida hylätä mahdollisena jakaumana. Aina sopivaa jakaumaa ei välttämättä löydy heti, jolloin voi joutua miettimään alkuperäistä jakaumajoukkoa uudestaan. Toisaalta on myös mahdollista, että aineisto ei sovi mihinkään tunnettuun jakaumaan kovin hyvin.

Yleisesti ottaen tutkimusprosessi jakauman valitsemiseksi voidaan esittää lyhyesti seuraavasti:

1. Visuaalinen diagnostiikka ja perussuureiden, kuten keskiarvon ja mediaanin laskeminen.
2. Mallin rajaaminen erilaisten kuvaajien avulla. Tässä työssä käytetään QQ-kuvaajia, riskifunktioiden kuvaajia ja odotettu ylitys -funktion kuvaajia.
3. Jatkotutkimus parhaiten sopivimmille malleille tilastollisilla keinoilla.

Vaihe yksi alkaa havainnollistamalla aineistoa kuvassa 5.1, ja esittelemällä kvantiileja sekä aineistosta että eksponenttijakaumasta. Vaiheen yksi tutkimuksesta saadaan merkittävä syy epäillä, että jakauma on paksuhäntäinen. Vaihe kaksi aloitetaan esittelemällä paksuhäntäisten jakaumien teoriaa, joka on olennaista tutkimuksen kannalta. Tämän jälkeen esitellään riskifunktio ja odotetun ylityksen funktio, joiden avulla verrataan laskujen jakaumaa erilaisiin paksuhäntäisiin jakaumiin. Todetaan, että laskut saattaisivat olla otos Pareto-jakaumasta. Vai-



Kuva 5.1: Pylväskuvaaja satunnaisesti valituista 1000 laskusta

heessa kolme tarkastelemme lähemmin Pareto-jakauman ominaisuuksia ja mitä se tarkoittaa laskujen muodostumisen kannalta.

Tutkimuksen tavoitteena on löytää todennäköisyysjakauma, joka olisi voinut tuottaa havaitun otoksen. Tutkimusmenetelmän lähtökohtana on oletus siitä, että havaitut arvot ovat samasta jakaumasta, ja että ne ovat toisistaan riippumattomia havaintoja. Todellisuudessa jotkin potilaat eivät välttämättä ole riippumattomia, esimerkiksi jos auto-onnettomuus aiheuttaa useamman henkilön vahingon. Lähtökohtaisesti voidaan kuitenkin olettaa, että yksittäisen henkilön kohdalla sairaalahoidon tarve on riippumaton muista henkilöistä. Edellä mainitun auto-onnettomuuden voidaan esimerkiksi ajatella olevan kunkin henkilön kohdalla tapahtuneet erilliset vahingot.

Tutkimus alkaa potilaiden kustannusten analysoimisesta. Yksinkertaisilla tilastollisilla menetelmillä, mediaanilla ja keskiarvolla, huomataan, että keskiarvo on huomattavasti mediaania suurempi. Aineiston mediaani on 1327 euroa, kun taas keskiarvo on 3704 euroa. Tämä suuri

ero kertoo siitä, että aineistossa on paljon pieniä havaintoja, mutta lukumääräisesti vähän suuria havaintoja. Kuvasta 5.1 nähdään, että suurin osa laskuista on matalalla tasolla, mutta osa laskuista on huomattavasti korkeammalla tasolla.

Tämä laskujen epätasaisuus selittää keskiarvon suuremman tason verrattuna mediaaniin. Muutama iso havainto riittää nostamaan keskiarvoa huomattavasti, mikäli iso havainto on suhteessa moninkertainen pieneen havaintoon. Aineistosta voidaan määritellä pienet ja isot havainnot myös matemaattisesti empiirisen kvantiilifunktion avulla.

Taulukko 5.1: Potilaiden laskujen kvantiilien arvot.

25%	50%	75%	95%	99%
404	1327	3983	13720	33250

Taulukko 5.2: Eksponenttijakauman kvantiilit ($\lambda = 1/1915$).

25%	50%	75%	95%	99%
550	1327	2654	5736	8818

Taulukosta 5.1 huomataan, että laskujen 75-kvantiili on huomattavasti kauempana mediaanista kuin 25-kvantiili. Lisäksi havaitaan, että laskujen isommat kvantiilit kasvavat huomattavasti suuremmiksi. Vertailun vuoksi taulukkoon 5.2 on laskettu parametrilla 1/1915 eksponenttijakauman arvot. Parametri valittiin siten, että 50-kvantiili, eli mediaani, olisi yhtä suuri kuin potilaiden laskujen tapauksessa. Näitä kahta kvantiilitaulukkoa katsomalla huomataan, että laskujen ylemmät kvantiilit ovat selvästi korkeampia kuin eksponenttijakaumalla. Tämä viittaa siihen, että laskujen jakauma olisi jossakin mielessä paksuhäntäinen. Tämä termi määritellään tarkemmin alla määritelmässä 5.1.1.

5.1 Paksuhäntäiset jakaumat

Määritellään ensiksi paksuhäntäinen satunnaismuuttuja. Kun satunnaismuuttuja on määritelty, voidaan aloittaa potilaiden laskuihin sopivan jakauman hahmottelu. Tästä eteenpäin merkintä log tarkoittaa luonnollista logaritmia.

Määritelmä 5.1.1. *Satunnaismuuttuja X on paksuhäntäinen oikealta puolelta, jos kaikilla $s > 0$ pätee*

$$\mathbb{E}(e^{sX}) = \infty.$$

Jos satunnaismuuttuja on paksuhäntäinen, sillä ei ole momenttiemäfunktiota. Tämän työn yhteydessä paksuhäntäisyydellä tarkoitetaan nimenomaan paksuhäntäisyyttä oikealta puolelta, sillä kustannukset ovat ei-negatiivisia. Paksuhäntäisyys tarkoittaa käytännössä sitä, että huomattavan suurella todennäköisyydellä jakaumasta satunnaisesti otettu arvo on iso. Kaikkein paksuhäntäisimmillä satunnaismuuttujilla ei ole edes odotusarvoa, kuten esimerkiksi Cauchy-jakautuneella satunnaismuuttujalla. Tällainen jakauma ei kuitenkaan kovin todennäköisesti kuvaa potilaiden kustannuksia joten tarkastelemme hieman kevyempihäntäisiä, mutta määritelmän 5.1.1 mielessä paksuhäntäisiä muuttujia.

Paksuhäntäisyyden voi esittää myös häntäjakauman avulla, kuten seuraavassa lauseessa osoitetaan. Jatkossa oletetaan, että satunnaismuuttujat ovat ei-negatiivisia.

Lause 5.1. *Satunnaismuuttuja X on paksuhäntäinen, jos ja vain jos*

$$\limsup_{x\to\infty} \frac{\log(\mathbb{P}(X > x))}{x} = 0 \tag{5.1}$$

Todistus. Oletetaan ensiksi, että satunnaismuuttuja on paksuhäntäinen ja osoitetaan, että yhtälö 5.1 pätee. Etsitään raja-arvolla sopivat ylä- ja alaraja. Ensiksi voidaan approksimoida raja-arvoa yläpuolelta:

$$\limsup_{x\to\infty} \frac{\log(\mathbb{P}(X > x))}{x} \leq \limsup_{x\to\infty} \frac{\log(1)}{x} = 0.$$

Seuraavaksi etsitään raja-arvolle alaraja. Käytetään apuna yleisesti tunnettua kaava ei-negatiiviselle satunnaismuuttujalle

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x)\mathrm{d}x. \tag{5.2}$$

Paksuhäntäisen jakauman määritelmän, $\mathbb{E}(e^{sX}) = \infty$, kaikilla $s > 0$, ja kaavan 5.2 mukaan saadaan

$$\mathbb{E}\left(e^{sX}\right) = \int_0^\infty \mathbb{P}(e^{sX} > x)\mathrm{d}x = \infty, \quad \text{kaikilla}\, s > 0.$$

Koska integraali on ääretön, tiedetään että on olemassa kasvava jono $(x_n) \uparrow \infty$, jolle pätee, että

$$\mathbb{P}(e^{sX} > x_i) \geq \frac{1}{x_i^{1+\varepsilon}}, \quad \text{kaikilla}\, i, \,\text{kun}\, \varepsilon > 0.$$

Huomataan, että

$$\mathbb{P}(e^{sX} > x_i) \geq \frac{1}{x_i^{1+\varepsilon}} \iff \mathbb{P}\left(X > \frac{\log(x_i)}{s}\right) \geq \frac{1}{x_i^{1+\varepsilon}}.$$

Sijoitetaan jälkimmäiseen yhtälöön $x_i = e^{y_i s}$, jolloin saadaan

$$\mathbb{P}\left(X > y_i\right) \geq \frac{1}{e^{(1+\varepsilon)y_i s}}.$$

Nyt voidaan arvioida yhtälön 5.1 raja-arvoa alapuolelta, jolloin saadaan

$$\limsup_{x\to\infty} \frac{\log(\mathbb{P}(X > x))}{x} \geq \limsup_{i\to\infty} \frac{\log(e^{-(1+\varepsilon)y_i s})}{e^{y_i s}} = \limsup_{i\to\infty} \frac{-(1+\varepsilon)y_i s}{e^{y_i s}} = 0.$$

Koska ε voidaan valita mielenvaltaisen pieneksi, tämä todistaa toisen suunnan lauseesta.

Seuraavaksi oletetaan, että yhtälö 5.1 pätee. Todistetaan, että X on paksuhäntäinen osoittamalla ristiriita oletuksesta, että X ei ole paksuhäntäinen. Oletetaan, että on olemassa $s > 0$, jolle pätee $\mathbb{E}(e^{sX}) < \infty$. Odotusarvoa voidaan arvioida seuraavasti

$$\mathbb{E}(e^{sX}) \geq \mathbb{E}(e^{sX}\mathbb{I}(X > x)) \geq \mathbb{E}(e^{sx}\mathbb{I}(X > x)) = e^{sx}\mathbb{P}(X > x).$$

Tämän avulla saadaan

$$\mathbb{P}(X > x) \leq \mathbb{E}(e^{sX})e^{-sx}.$$

Tämä voidaan sijoittaa yhtälön 5.1 raja-arvoon, jolloin saadaan

$$\limsup_{x\to\infty} \frac{\log(\mathbb{P}(X > x))}{x} \leq \limsup_{x\to\infty} \frac{\log(\mathbb{E}(e^{sX})e^{-sx})}{x} = \limsup_{x\to\infty} \frac{-sx}{x} = -s.$$

Mikä johtaa ristiriitaan, sillä raja-arvo ei voi olla samaan aikaan sekä negatiivista lukua pienempi että nolla. □

Yksi tapa mitata satunnaismuuttujan paksuhäntäisyyttä on tutkia sen kertymäfunktion häntäfunktion vähenemisnopeutta. Tähän voidaan käyttää hyväksi momentti-indeksin muotoilua.

Määritelmä 5.1.2. *Satunnaismuuttujalla X on potenssihäntä indeksillä α, jos*

$$\lim_{x\to\infty} \frac{\log(\mathbb{P}(X > x))}{\log x} = -\alpha. \tag{5.3}$$

Tämän tarkoitus voidaan esittää myös toisessa muodossa. Oletetaan, että satunnaismuuttujalla X on kertymäfunktio $F(x)$ ja $\bar{F}(x) = 1 - F(x)$ on sen häntäfunktio, ja että sille pätee ehto 5.3 arvolla α. Tällöin häntäfunktiolle pätee suunnilleen, että $F(x) \approx x^{-\alpha}$. Raja-arvon määritelmän mukaan jokaiselle $\varepsilon > 0$ on olemassa sellainen x', että

$$\left|\frac{\log(\mathbb{P}(X > x_\varepsilon))}{\log x_\varepsilon} + \alpha\right| < \varepsilon,$$

kun $x_\varepsilon > x'$. Tästä voidaan itseisarvon määritelmän avulla ratkaista ylä- ja alaraja todennäköisyydelle $\mathbb{P}(X > x)$ seuraavasti:

$$-\varepsilon < \frac{\log(\mathbb{P}(X > x_\varepsilon))}{\log x_\varepsilon} + \alpha < \varepsilon \iff x_\varepsilon^{-\varepsilon-\alpha} < \mathbb{P}(X > x_\varepsilon) < x_\varepsilon^{\varepsilon-\alpha}.$$

Laskuissa voidaan olettaa, että $\log x_\varepsilon > 0$. Näistä epäyhtälöistä nähdään, että satunnaismuuttujan häntäfunktio on lähellä funktiota $x^{-\alpha}$, mikäli sillä on potenssihäntä indeksillä α.

Esitellään seuraavaksi kolme erilaista yleisesti tunnettua paksuhäntäistä jakaumaa.

Esimerkki 5.1.3. Pareto-jakauma on tunnettu paksuhäntäinen jakauma, ja se on paksuhäntäisin tässä esiteltävistä jakaumista. Pareto-jakauman kertymäfunktio on:

$$F_{Pareto}(x) = \begin{cases} 1 - \left(\frac{x}{b}\right)^{-\alpha} & x \geq b \\ 0, & x < b. \end{cases}$$

Parametri α on jakauman indeksi ja b on jakauman skaalaparametri. Tässä tutkimuksessa, jos parametrejä ei erikseen kerrota, oletetaan molempien olevan 1. Pareto-jakauman momentit ovat olemassa korkeintaan indeksin itseisarvoon asti. Tämä voidaan nähdä käyttämällä kaavaa 5.2, josta saadaan

$$\mathbb{E}(X^n) = \int_0^\infty \mathbb{P}(X^n > x)\mathrm{d}x = \int_0^\infty \mathbb{P}(X > x^{\frac{1}{n}})\mathrm{d}x = b^n + \int_{b^n}^\infty \left(\frac{x^{\frac{1}{n}}}{b}\right)^{-\alpha} \mathrm{d}x = b^n + \int_{b^n}^\infty b^\alpha \frac{1}{x^{\frac{\alpha}{n}}}\mathrm{d}x,$$

missä X on Pareto-jakautunut satunnaismuuttuja. Jakauman määritelmän mukaan $\bar{F}(x^{\frac{1}{n}}) = 1$, kun $x < b^n$. Huomataan, että oikean puoleisin integraali ei suppene, jos $n \geq \alpha$. Esimerkiksi, jos $\alpha < 1$, niin jakaumalla ei ole odotusarvoa.

Esimerkki 5.1.4. Log-normaalijakauma on hieman Pareto-jakaumaa kevythäntäisempi. Log-normaali satunnaismuuttuja saadaan standardinormaalijakaumasta. Jos Z on standardinormaalijakautunut satunnaismuuttuja, niin satunnaismuuttujalla $X = e^{\mu+\sigma Z}$ on log-normaalijakauma. Log-normaalijakauman kertymäfunktio on muotoa:

$$F_{LN}(x) = \Phi\left(\frac{(\log x) - \mu}{\sigma}\right),$$

missä Φ on standardinormaalijakauman kertymäfunktio. Parametrit μ ja σ kertovat satunnaismuuttujan logaritmin odotusarvon ja varianssin. Oletusarvoisesti log-normaalin jakauman parametrit ovat $\mu = 0$ ja $\sigma = 1$.

Esimerkki 5.1.5. Weibull-jakauma on tässä esitellyistä paksuhäntäisistä jakaumista kevythäntäisin. Weibull-jakauman kertymäfunktio on muotoa:

$$F_{Weibull}(x) = \begin{cases} 1 - e^{-(x/\lambda)^k}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Parametrit k ja λ ovat nimiltään muoto ja skaala. Weibull-jakauma on paksuhäntäinen ainoastaan jos $k < 1$. Weibull-jakauma on myös yleistys eksponenttijakaumasta, sillä parametrillä $k = 1$, se on eksponenttijakauma parametrillä λ^{-1}.

Seuraavaksi näytetään, miten kevythäntäisistä satunnaismuuttujista voi syntyä paksuhäntäinen muuttuja tulon kautta.

Esimerkki 5.1.6. Oletetaan, että X_1 ja X_2 ovat riippumattomia eksponentiaalisesti jakautuneita satunnaismuuttujia parametrilla λ. Eksponentiaaliset satunnaismuuttujat ovat kevythäntäisiä. Muodostetaan uusi satunnaismuuttuja $X = X_1 \cdot X_2$. Lauseesta 5.1 tiedetään, että jos

$$\limsup_{x\to\infty} \frac{\log(\mathbb{P}(X > x)}{x} = 0, \tag{5.4}$$

niin satunnaismuuttuja X on paksuhäntäinen. Todetaan ensiksi, että

$$\{X_1 > \sqrt{x}, X_2 > \sqrt{x}\} \subseteq \{X_1 X_2 > x\}.$$

Nyt voidaan laskea

$$\limsup_{x\to\infty} \frac{\log(\mathbb{P}(X > x))}{x} = \limsup_{x\to\infty} \frac{\log(\mathbb{P}(X_1 X_2 > x))}{x} \geq \limsup_{x\to\infty} \frac{\log(\mathbb{P}(X_1 > \sqrt{x})\mathbb{P}(X_2 > \sqrt{x}))}{x}$$

$$= \limsup_{x\to\infty} \frac{\log(e^{-\lambda\sqrt{x}}) + \log(e^{-\lambda\sqrt{x}})}{x} = \limsup_{x\to\infty} \frac{-2\lambda\sqrt{x}}{x} = 0.$$

Yhtälön 5.4 raja-arvon alarajaksi saatiin 0. Koska todennäköisyyttä voidaan approksimoida ylhäältä päin arvolla 1, saadaan raja-arvolle ylärajaksi arvo 0. Täten satunnaismuuttuja X on paksuhäntäinen.

5.1.1 Riskifunktio

Riskifunktio, englanniksi hazard function, on deterministinen muunnos häntäfunktiosta. Sen avulla on yksinkertaista havainnollistaa häntäfunktion vähenemisvauhtia, koska paksut hännät vastaavat hitaasti kasvavia riskifunktioita ja kevyet hännät nopeasti kasvavia riskifunktioita. Jokainen kasvava funktio, jolla $R(0) = 0$, voidaan tulkita jonkin satunnaismuuttujan riskifunktioksi. Aloitetaan määrittelemällä riskifunktio, ja esitellään kuvaajia joissa on otoksesta laskettu riskifunktio, sekä teoreettisten jakaumien riskifunktioita.

Määritelmä 5.1.7. *Satunnaismuuttujan X riskifunktio $R(x)$ on*

$$R(x) = -\log(\mathbb{P}(X > x)).$$

Otoksen tapauksessa käytetään empiiristä häntäfunktiota, joka on muunnos empiirisestä kertymäfunktiosta. Empiirisestä häntäfunktiosta lasketaan empiirisen riskifunktion arvot. Ensiksi huomataan, että eksponenttijakauman riskifunktio on:

$$R_{Exp}(x) = -\log(e^{-\lambda x}) = \lambda x.$$

Eksponenttijakauman riskifunktio on siis lineaarinen. Tämä toimii eräänlaisena raja-arvona jakauman paksuhäntäisyydelle, sillä paksuhäntäisten jakaumien riskifunktiot ovat lineaarista hitaammin kasvavia funktioita. Esimerkiksi Pareto-jakauman indeksillä 1 riskifunktio on

$$R_{Pareto} = -\log\left(\left(\frac{1}{x}\right)^1\right) = \log(x).$$

Seuraavaksi tarkastellaan miltä jakaumasta laskettu empiirinen riskifunktio näyttää, ja verrataan sitä erilaisten teoreettisten jakaumien riskifunktioihin. Kuvaajaan 5.2a on piirretty neljän erilaisen jakauman ja otoksen riskifunktiot. Kuvaajassa on korostettu eksponenttijakauman riskifunktion lineaarisuutta, piirtämällä kuvaajan akselit samanmittaisina. Kuvaajasta huomataan selvästi, että paksuhäntäiset jakaumat ovat selvästi tämän suoran alla, varsinkin kun poistutaan nollan ympäristöstä. Huomataan, että otos sijoittuu Pareto- ja log-normaalijakauman väliin, ja voidaan päätellä, että potilaiden laskujen jakauma on paksuhäntäinen.

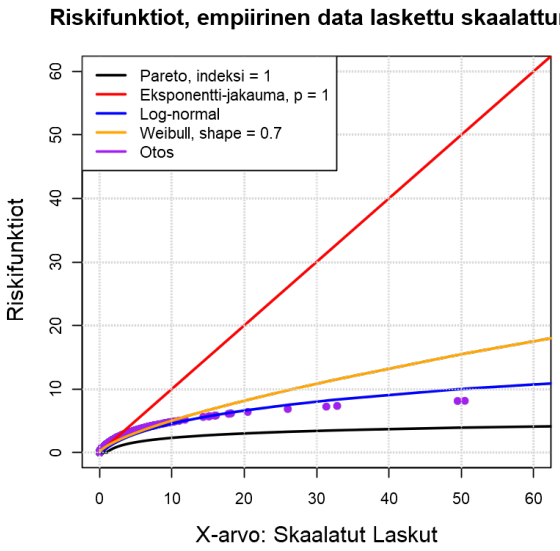
Kuvaajassa 5.2b tutkitaan tarkemmin näitä paksuhäntäisten jakaumien riskifunktioita, ja nähdään selvemmin riskifunktioiden kuvaajien erot. Otoksen riskifunktion kuvaaja on lähimpänä log-normaalin riskifunktioin kuvaajaa. Kuvaajasta voidaan kuitenkin huomata, että otoksen riskifunktion kasvunopeus näyttäisi olevan hitaampi kuin log-normaali-jakaumalla. Tämä viittaisi siihen, että otoksen jakauma olisi paksuhäntäisempää kuin log-normaali-jakauma.

Seuraavaksi vertaillaan log-normaali- ja Pareto-jakauman riskifunktioita erilaisilla parametrien arvoilla. Nämä parametrit vaikuttavat lähinnä jakauman paksu-häntäisyyteen. Pareto-jakaumassa parametrin arvo kertoo suoraan jakauman indeksin. Log-normaali-jakaumassa vaihdetaan jakauman generoivan normaalijakauman varianssin arvoa, minkä suurentaminen tekee jakaumasta paksuhäntäisemmän.

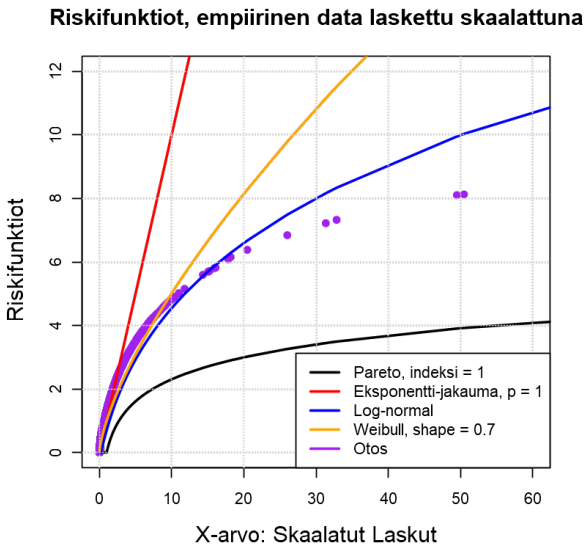
Kuvaajassa 5.3a on laskettuna erilaisten log-normaali-jakaumien riskifunktioita, ja kuvaajassa 5.3b taas on laskettu erilaisten Pareto-jakaumien riskifunktioita. Näistä kuvista voidaan päätellä tarkemmin, mikä jakauma kuvaa parhaiten potilaiden laskujen jakautumista. Kuvasta 5.3b nähdään selvästi, että indeksillä 2 Pareto-jakauma on hyvin lähellä otoksesta laskettua riskifunktiota. Tämä on yksi perustelu Pareto-jakauman valitsemiseen mallintamisessa, mutta varmempi tulos saadaan, kun tutkitaan asiaa myös muilla lähestymistavoilla.

5.1.2 QQ-kuvaaja

Kun halutaan tutkia mikä jakauma jollain havaitulla otoksella saattaisi olla, niin yksi tekniikka on kvantiili-kvantiili-kuvaajien, eli lyhyemmin QQ-kuvaajien, käyttö. QQ-kuvaajassa verrataan havaittua otosta johonkin teoreettiseen jakaumaan tai toiseen otokseen, käyttämällä hyväksi näiden kvantiileja. QQ-kuvaajassa laitetaan otoksen ja siihen verrattavan jakauman vastaavat kvantiilit koordinaatistoon. Mikäli jakaumat ovat identtiset, pisteiden tulisi asettua suoralle $x = y$. Tämä on vaikeaa saavuttaa todellisilla otoksilla, sillä se vaatisi jakauman tuntemis-



(a) Eri jakaumien ja otoksen riskifunktioita, piirrettynä tasasivuiseen kuvaajaan. Otoksen riskifunktio on selvästi suoran $x = y$ alapuolella.



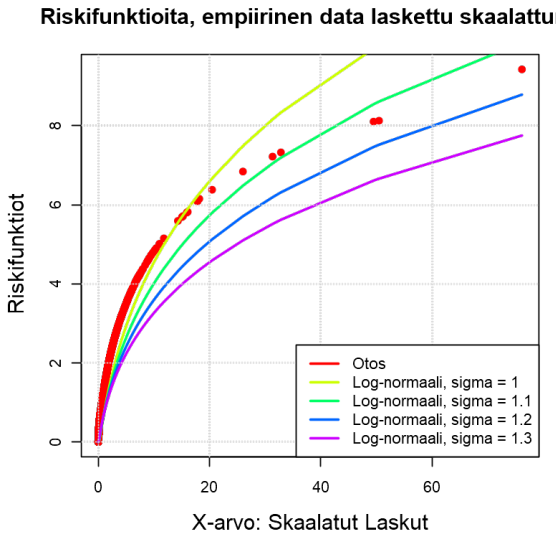
(b) Riskifunktioita eri jakaumista ja otoksesta, tarkennettuna paksuhäntäisiin jakaumiin. Otoksen riskifunktio on log-normaali-jakauman alla.

Kuva 5.2: Kaksi kuvaajaa riskifunktioista. Vasemmassa akselit ovat samanmittaiset, oikeassa taas korostettu paksuhäntäisiä jakaumia.

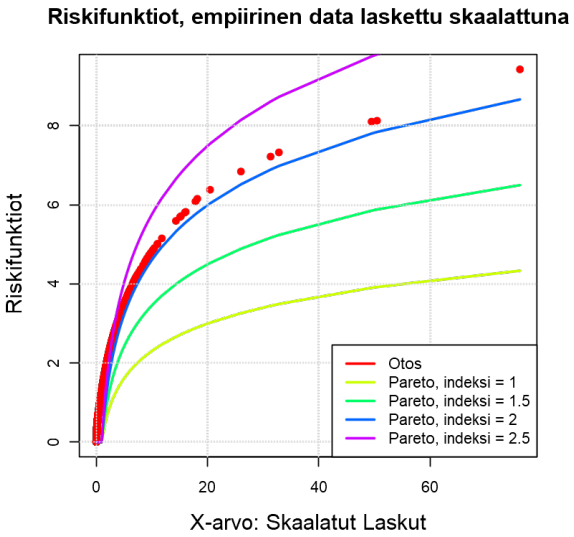
ta etukäteen. Jos verrattava jakauma on kuitenkin lineaarinen transformaatio otoksesta, niin silloin pisteiden pitäisi asettua jollekin suoralle.

Lisäksi pisteiden sijainnista voi päätellä onko otoksen jakauma paksu- vai kevythäntäisempi kuin verrattavan. QQ-kuvaajiin voidaan käyttää erilaisia tunnettuja jakaumia, ja tulkita kuvaajista mikä jakauma sopii parhaiten havaittuun otokseen. Kuva A.9 esittelee kuusi erilaista QQ-kuvaajaa, jossa verrataan sekä eksponenttijakaumaa että laskuja erilaisiin jakaumiin. Mustat pisteet ovat kvantiilien pisteitä. Kuvaajiin on myös piirretty sinisellä viiva, jossa pisteiden tulisi olla jotta ne olisivat verrattavan jakauman mukaisia. Sinisellä katkoviivalla on havainnollistettu 95-prosentin luottamusväliä kvantiilien arvoissa.

Kuvasta A.9a huomataan, että jos verrataan kahta samanlaista jakaumaa, pisteet asettuvat hyvin lähelle $x = y$ viivaa, vaikka kuitenkaan kaikki eivät välttämättä ole juuri kyseisellä suoralla. Kuvaajasta A.9d voisi päätellä, että log-normaalijakauma olisi todennäköisesti paras jakauma kuvaamaan laskuja. Kuvissa A.9e ja A.9c näkyvät Weibull ja eksponentiaalijakaumat ovat selvästi liian kevythäntäisiä verrattuna laskuihin, sillä pisteet asettuvat selvästi suoran yläpuolelle. Kuvassa A.9f havaitaan, että jakauma ei ole lähellä Pareto-jakaumaa indeksillä 1.



(a) Eri log-normaali-jakauman riskifunktioita verrattuna otoksen riskifunktioon.



(b) Pareto-jakauman riskifunktioita eri indeksin arvoilla verrattuna otoksen riskifunktioon.

Kuva 5.3: Erilaisia log-normaali- ja Pareto-jakaumia verrattuna otoksen riskifunktioihin. Huomataan, että Pareto-jakauma vastaa paremmin otoksen riskifunktioita.

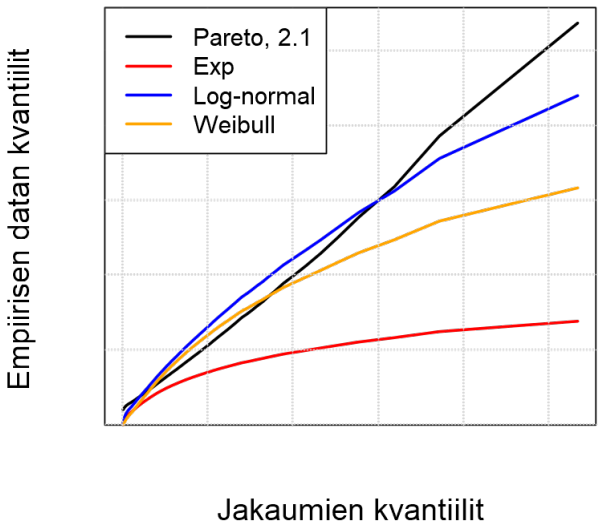
Palautetaan mieleen kuitenkin edellisessä kohdassa saavutettu tulos, että jakauma voisi olla Pareto-jakautunut hieman kahta suuremmalla indeksillä. Piirretään siis QQ-kuvaaja Pareto-jakauman indeksillä 2, ja tutkitaan mikä on lopputulos. Kuvaaja on piirrettynä kuvaan 5.5a, josta nähdään, että pisteet ovat asettuneet suoralle, joka on hieman $x = y$ suoraa loivempi. Tämän jälkeen voidaan kokeilla miltä näyttäisi kuvaaja näyttäisi, jos indeksi olisi hieman isompi, esimerkiksi 2.1. Tämä kuvaaja on esiteltynä kuvassa 5.5b. Tästä kuvaajasta nähdään, että pisteet ovat asettuneet lähes täydellisesti suoralle.

Kuvasta 5.4 huomataan, että Pareto-jakauman indeksillä 2.1 kvantiilifunktion pisteet asettuvat lähes suoralle, kun niitä verrataan empiiriseen kvantiilifunktioon. Vastaavasti log-normaali-jakauman kvantiilifunktio taittuu tämän suoran alle, mikä viittaisi siihen, että log-normaali-jakauma ei ole niin paksuhäntäinen kuin potilaiden laskujen jakauma.

Muut kuvassa 5.4 esitellyt jakaumat ovat selvästi suoran alapuolella, ja siten liian kevythäntäisiä verrattuna laskuihin.

5.1.3 Odotetun ylityksen -kuvaaja

Paksuhäntäisistä muuttujista halutaan mahdollisesti tietää, kuinka suuri pienen todennäköisyyden tapahtuma on. Sitä voidaan arvioida käyttämällä odotettua ylitystä (Mean excess). Odo-



Kuva 5.4: Eri jakaumien kvantiilifunktioita verrattuna laskujen kvantiileihin, piirrettyinä samaan kuvaajaan.

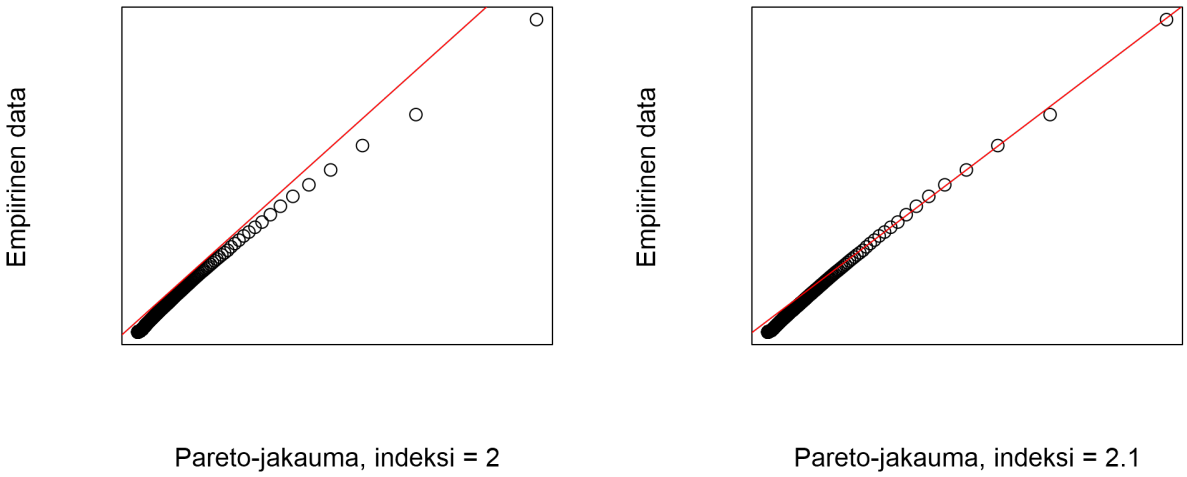
tetulla ylityksellä tarkoitetaan keskiarvoa siitä, minkä verran satunnaismuuttuja ylittää jonkin tietyn raja-arvon. Odotetun ylityksen teoriaa on esitelty tarkemmin artikkelissa [7]. Odotettu ylitys -funktiossa funktio ottaa lähtöarvoksi raja-arvon, ja palauttaa arvona sen arvon, minkä verran satunnaismuuttuja keskimäärin ylittää tämän raja-arvon. Tarkemmin se voidaan määritellä ehdollisen odotusarvon avulla.

Määritelmä 5.1.8. *Satunnaismuuttujan X odotettu ylitys funktio $M(u)$ on*

$$M(u) = \mathbb{E}(X - u | X > u).$$

Odotettu ylitys funktiota voidaan käyttää arvioimaan satunnaismuuttujan ominaisuuksia, kuten paksuhäntäisyyttä. Kevyt-häntäisillä satunnaismuuttujilla odotettu ylitys funktio pienenee tai pysyy vakiona kun $x \to \infty$. Pareto-jakaumalla tämä kuvaaja on lineaarisesti kasvava. Kuvassa 5.6b on esitelty erilaisten jakaumien odotetun ylityksen funktioita. Kuvaajasta huomataan, että kevythäntäiset jakaumat, normaalijakauma ja eksponenttijakauma, saavat paljon pienempiä arvoja kuin paksuhäntäiset jakaumat. Kuvaajasta huomataan myös Pareto-jakauman odotetun ylityksen lineaarisuus. Log-normaali -jakauman kuvaaja näyttää kasvavan hitaammin kuin Pareto-jakauman, mikä kertoo sen olevan kevythäntäisempi jakauma kuin Pareto-jakauma, ainakin kyseisillä parametreillä.

Kuvassa 5.6a on esitelty laskuista lasketun odotetun ylityksen funktion arvoja. Pisteistä



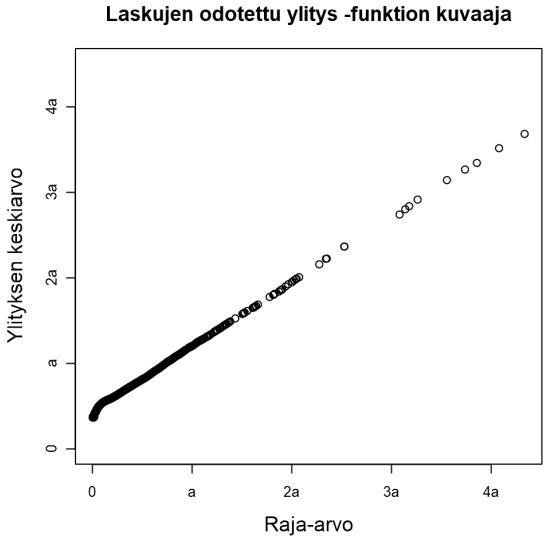
(a) QQ-kuvaaja Pareto-jakauman indeksillä 2 (b) QQ-kuvaaja Pareto-jakauman indeksillä 2.1

Kuva 5.5: QQ-kuvaajat kahdesta Pareto-jakaumasta. Punainen viiva on $x = y$ suora. Huomataan että vasemman puoleisessa pisteet ovat asettuneet lähes täydellisesti suoralle.

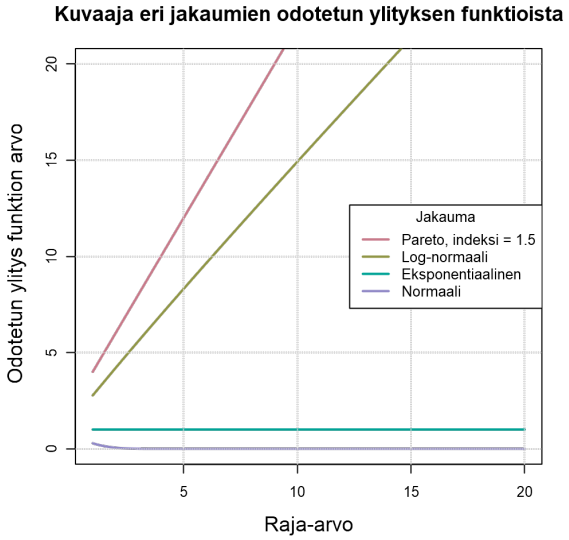
huomataan, että ne asettuvat selvästi suoralle tietyn arvon jälkeen, mikä on ominaista Pareto-jakaumalle. Tämä on yksi osoitus siitä, että laskujen jakauma olisi Pareto-jakautunut ainakin suurilla arvoilla.

5.2 Jakauman häntäparametrin tutkiminen

Voidaan perustellusti siis todeta, että jakauma on paksuhäntäinen, ja todennäköisesti suurimmat arvot noudattavat Pareto-jakaumaa. Täten voidaan tutkia laskujen jakaumaa tarkemmin käyttämällä hyväksi menetelmiä, joissa oletetaan jakauman kuuluvan Pareto-jakaumaperheeseen jonkin tietyn rajan jälkeen. Paksuhäntäisen jakauman häntäfunktion indeksin estimointi on tärkeää, kun halutaan oppia tuntemaan jakauman ominaisuuksia. Häntäfunktion indeksi kertoo, kuinka nopeasti suurten tapahtumien todennäköisyys menee nollaan. Indeksin tunteminen antaa selkeän kuvan suurten tapahtumien todennäköisyydestä, ja auttaa varautumaan paremmin harvinaisiin mutta kalliisiin tapauksiin. Tässä osiossa esitellään erilaisia tapoja estimoida indeksiä, sekä perustellaan sekä Pareto-jakauman käyttö estimoinnissa, että indeksin estimoinnissa käytettävät oletukset jakaumasta.



(a) Laskuista laskettu odotettu ylitys -funktio. Akseleiden arvoissa a on tunnettu vakio.



(b) Erilaisten jakaumien odotetun ylityksen funktioita.

Kuva 5.6: Kuvaajat kahdesta odotetun ylityksen funktioista. Vasemmalla laskuista laskettu empiirinen odotettu ylitys, ja oikealla erilaisia teoreettisista jakaumista laskettuja odotettuja ylityksiä. Huomataan, että kevythäntäisten jakaumien funktiot saavat huomattavasti pienempiä arvoja.

5.2.1 Estimoinnin teoriaa

Paksuhäntäisen jakauman indeksin estimointi on kriittinen osa jakauman estimointia, josta on julkaistu lukuisia artikkeleita. Eräs aikaisimmista estimaateista jakauman indeksille oli Hillin estimaatti [8]. Hillin estimaatti olettaa, että havaitun otoksen kertymäfunktio on muotoa $F(x) = 1 - Cx^{-\alpha}$, kun $x > u$, missä u on jokin tunnettu raja. Tämä tarkoittaa sitä, että Hillin estimaatti olettaa otoksen olevan Pareto-jakautunut tietyn rajan jälkeen. Estimaatissa käsitellään havaintoja järjestettynä jonona, joten merkitään, että $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(k)}$ ovat k isointa havaintoa ja $X_{(k+1)} > u$. Hillin estimaatin $\hat{\alpha}$ kaava on:

$$\hat{\alpha}_k^{-1} = \frac{1}{k}\left(\sum_{i=1}^{k} \log\left(\frac{X_{(i)}}{X_{(k+1)}}\right)\right).$$

Estimaatin parametri k kertoo, kuinka montaa suurimmista arvoista käytetään estimaatin laskemisessa.

Esitetään seuraavaksi mihin Hillin estimaatti perustuu. Lähtöoletus on se, että arvot ovat Pareto-jakautuneita, ja summan sisällä tarkastellaan suhdetta, jossa katsotaan kuinka paljon

suurempia havainnot $1, \ldots, k$ ovat verrattuna havaintoon $k+1$. Tämä suhde on vielä logaritmoitu, jonka jälkeen summa jaetaan havaintojen lukumäärällä. Kyseessä on siis logaritmoitu etäisyyden keskiarvo arvosta $X_{(k+1)}$.

Hillin estimaatin oletuksesta seuraa, että häntäfunktio on muotoa $\bar{F}(x) = Cx^{-\alpha}$. Tällöin $\log\left(\frac{X_{(i)}}{X_{(k+1)}}\right)$ on jakautunut seuraavasti:

$$\mathbb{P}\left(\log\left(\frac{X_{(i)}}{X_{(k+1)}}\right) > x\right) = \mathbb{P}\left(\frac{X_{(i)}}{X_{(k+1)}} > e^x\right) = \mathbb{P}\left(X_{(i)} > e^x X_{(k+1)}\right) = \bar{F}\left(e^x X_{(k+1)}\right)$$

$$= C\left(e^x X_{(k+1)}\right)^{-\alpha} = \frac{C}{X_{(k+1)}^{\alpha}} e^{-\alpha x}.$$

Mistä havaitaan, että estimaatissa esiintyvä termi $\log\left(\frac{X_{(i)}}{X_{(k+1)}}\right)$ on lineaarinen muunnos eksponenttijakaumasta parametrilla α.

Jos oletettaisiin lisäksi, että havaittu otos olisi Pareto-jakautunut kaikilla positiivisilla arvoilla, eikä vasta jonkun raja-arvon jälkeen, ja että $C = 1$, niin tällöin voidaan valita k siten, että:

$$\frac{C}{X_{(k+1)}^{\alpha}} e^{-\alpha x} = e^{-\alpha x}.$$

Tällöin $\log(X_{(i)})$ satunnaismuuttuja olisi siis eksponentiaalisesti jakautunut parametrin arvolla α, ja parametriä voidaan estimoida otoksen keskiarvon käänteisluvulla. Tästä yhteydestä saadaan perustelu Hillin estimaatille. Seuraavaksi esitellään tapoja estimoida indeksiä, jos tehdään Hillin estimaatin oletusta kevyempi oletus.

Monien estimaattoreiden lähtöoletus on se, että jakauman oletetaan olevan säännöllisesti vaihteleva jollain indeksillä. Määritellään ensiksi, mitä tämä tarkoittaa.

Määritelmä 5.2.1. *Funktio $f(x)$ on säännöllisesti vaihteleva indeksillä α jos kaikilla $t > 0$ pätee:*

$$\lim_{x\to\infty} \frac{f(tx)}{f(x)} = t^{\alpha}.$$

Tällöin voidaan merkitä $f(x) \in \mathcal{R}_{\alpha}$. Jos $\alpha = 0$, sanotaan, että funktio on hitaasti vaihteleva.

Tämä ominaisuus tarkoittaa sitä, että funktio käyttäytyy tunnetulla tavalla kun x kasvaa suureksi. Tällainen ominaisuus tekee funktioista helposti käsiteltävän, kun $x \to \infty$. Lisäksi monet tunnetut funktiot ovat säännöllisestä vaihtelevia, kuten esimerkiksi Pareto-jakauman häntäfunktio.

$$\lim_{x\to\infty} \frac{\bar{F}(tx)}{\bar{F}(x)} = \lim_{x\to\infty} \frac{(tx)^{-\alpha}}{x^{-\alpha}} = t^{-\alpha}.$$

Säännöllisesti vaihteleville funktioille on todistettu myös lukuisia hyödyllisiä ominaisuuksia, joista tässä työssä käytetään Karamatan esityslausetta. Sen avulla voidaan esittää kaikki

säännöllisesti vaihtelevat funktiot samassa muodossa, jota voidaan käyttää avuksi lauseiden todistamisessa.

Lause 5.2 (Karamatan esityslause)**.** *Jos funktio $f(x)$ on säännöllisesti vaihteleva indeksillä α, on olemassa $z \in (0, \infty)$, jolle pätee*

$$f(x) = c(x) \exp\left\{ \int_z^x \frac{a(t)}{t} \mathrm{d}t \right\}, \text{kaikilla } z < x < \infty.$$

Missä $c(x) \to c \in (0, \infty)$, ja $a(x) \to \alpha$, kun $x \to \infty$.

Lausetta ei todisteta tässä tutkimuksessa, vaan oletetaan tunnetuksi. Todistus löytyy säännöllistä vaihtelua käsittelevästä kirjasta [3] luvusta 1. Karamatan esityslauseesta voidaan johtaa myös muita muotoja säännöllisesti vaihtelevalle funktiolle.

Seuraus 5.1. *Jos funktio $f(x)$ on säännöllisesti vaihteleva indeksillä α, on olemassa sellainen $z \in (0, \infty)$, jolle pätee*

$$f_\alpha(x) = c(x) \exp\left\{ - \int_z^x \frac{1}{p(t)} \mathrm{d}t \right\},$$

missä $c(x) \to c \in (0, \infty)$ ja $\frac{t}{p(t)} \to -\alpha$, kun $t \to \infty$.

Todistus. Esityslauseen nojalla tiedetään, että funktiolle $f(x)$ on olemassa esitys:

$$f(x) = c(x) \exp\left\{ \int_z^x \frac{a(t)}{t} \mathrm{d}t \right\},$$

josta tiedetään, että $a(t) \to \alpha$, kun $t \to \infty$. Sijoitetaan $a(t) = -\frac{t}{p(t)}$, missä pätee $\frac{t}{p(t)} \to -\alpha$, kun $t \to \infty$. Nyt voidaan laskea:

$$f_\alpha(x) = c(x) \exp\left\{ \int_z^x \frac{a(t)}{t} \mathrm{d}t \right\} = c(x) \exp\left\{ \int_z^x \frac{-\frac{t}{p(t)}}{t} \mathrm{d}t \right\} = c(x) \exp\left\{ - \int_z^x \frac{1}{p(t)} \mathrm{d}t \right\}.$$

Esityslauseen nojalla tiedetään, että $c(x) \to c \in (0, \infty)$ ja $\frac{t}{p(t)} \to -\alpha$. □

Aiemmassa osiossa esitetyn odotetun ylityksen perusideaa voidaan käyttää myös suoraan jakauman tutkimiseen. Jos valitaan jakaumalle $F(x)$ raja u, jonka ylitystä halutaan tutkia, voidaan määritellä ylityksen suuruuden kertymäfunktio:

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u).$$

Tätä ylityksen suuruuksien tutkimista kutsutaan englanniksi peaks-over-threshold -metodiksi. Ylitysten tutkiminen mahdollistaa sen, että tutkimisessa voidaan keskittyä ainoastaan isoihin

arvoihin. Tällöin riittää olettaa, että ainoastaan suuret tapaukset olisivat jakautuneet tietyllä tavalla. Tämä on helpommin perusteltava oletus kuin se, että koko otos olisi jakautunut tietyllä tavalla. Isoimmat tapaukset ovat myös tutkimuksen kannalta mielenkiintoisimpia, sillä ne aiheuttavat merkittävimmät muutokset kustannuksissa.

Esitellään vielä yleistetty Pareto-jakauma, jota käytetään estimoinnissa. Yleistetyn Pareto-jakauman kertymäfunktio voidaan esittää seuraavasti:

$$G_{\alpha,\beta}(x) = \begin{cases} 1 - \left(1 + \frac{x}{\alpha\beta}\right)^{-\alpha} & \text{jos } \alpha \neq 0 \\ 1 - e^{\frac{-x}{\beta}} & \text{jos } \alpha = 0. \end{cases}$$

Pickands–Balkema–de Haan -teoreema antaa teoreettisen pohjan yleistetyn Pareto-jakauman käyttöön jakaumafunktion määrittelemisessä. Teoreeman pohjalta voidaan sanoa, että varsin yleisillä oletuksilla, jakauman odotettu ylitysfunktio lähestyy yleistettyä Pareto-jakaumaa. Lause on alunperin esitelty Pickandsin artikkelissa [16], sekä de Haanin ja Balkeman artikkelissa [2]. Lauseen perusteella voidaan etsiä yleistettyä Pareto-jakaumaa joka vastaa havaittua empiiristä jakaumaa tietyn arvon jälkeen.

5.2.2 Pickands-Balkema-de Haan -teoreema

Esitetään ensiksi lauseessa käytettyjä määritelmiä, ja lauseeseen läheisesti liittyviä tunnettuja lauseita. Tämän jälkeen esitetään itse lauseen yleinen tulos, jonka jälkeen todistetaan lauseesta tämän tutkielman kannalta oleellinen tapaus.

Määritellään tapa luokitella jakaumia samaan jakaumaperheeseen. Englanniksi termi on maximum domain of attraction, suomeksi jakauman vaikutuspiiri maksimin suhteen. Oletetaan, että on olemassa samoin jakautuneet ja riippumattomat satunnaismuuttujat $X_1, \ldots, X_n$, ja merkataan näiden maksimia $M_n = \max(X_1, \ldots, X_n)$. Nyt tarkoituksena on tutkia miten M_n käyttäytyy, kun $n \to \infty$.

Määritelmä 5.2.2. *Satunnaismuuttuja X, tai sen jakauma $F(x)$, kuuluu jakauman $G(x)$ vaikutuspiiriin maksimin suhteen, jos on olemassa sellaiset jonot (a_n) ja (b_n), että:*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \to G(x), \quad \text{kun } n \to \infty.$$

Määritelmä tarkoittaa siis, että jos satunnaismuuttujien maksimi on affiini muunnos jakaumasta $G(x)$, niin satunnaismuuttuja kuuluu sen vaikutuspiiriin maksimin suhteen. Vaikutuspiireille on olemassa ominaisuus, joka rajaa satunnaismuuttujan kuulumista eri vaikutuspiireihin. Seuraavaksi todistettavan lauseen nojalla, jos satunnaismuuttuja kuuluu jonkun tietyn jakauman vaikutuspiiriin, se voi kuuluu ainoastaan sen jakauman affiinien muunnoksien vaikutuspiireihin. Tätä lausetta kutsutaan yleisesti Convergence to types -teoreemaksi. Lauseessa

käytetään termiä ei-degeneroitunut jakauma, joka tarkoittaa tämän työn kannalta oleellisessa yksiulotteisessa tapauksessa sitä, että jakauman kertymäfunktio ei ole indikaattorifunktio. Todistuksessa mukaillaan Nyrhisen luentomonistetta [15].

Lause 5.3. *Oletetaan, että on olemassa lukujonot $a_n \in (0, \infty)$ ja $b_n \in \mathbb{R}$, siten että*

$$F^n(a_n x + b_n) \xrightarrow{d} G(x), \quad \text{kun } n \to \infty, \tag{5.5}$$

missä $G(x)$ on ei-degeneroitunut jakauma. Jos on olemassa lukujonot a'_n ja b'_n, siten että

$$F^n(a'_n x + b'_n) \xrightarrow{d} H(x), \quad \text{kun } n \to \infty, \tag{5.6}$$

missä $H(x)$ on ei-degeneroitunut jakauma, niin silloin pätee, että

$$\lim_{n\to\infty} \frac{a'_n}{a_n} = a \in (0, \infty) \quad \text{ja} \quad \lim_{n\to\infty} \frac{b'_n - b_n}{a_n} = b \in \mathbb{R}. \tag{5.7}$$

Lisäksi $H(x) = G(ax + b)$ kaikilla $x \in \mathbb{R}$. Väite pätee myös toiseen suuntaan, eli jos ehto 5.7 on voimassa, niin väite 5.6 pätee.

Todistus. Aloitetaan todistus olettamalla, että väite 5.6 pätee, ja osoitetaan väite 5.7 todeksi. Muunnetaan väitteen 5.6 funktio toiseen muotoon:

$$F^n(a'_n x + b'_n) = F^n\left(a_n\left(\frac{a'_n x}{a_n} + \frac{b'_n}{a_n}\right)\right) = F^n\left(a_n\left(\frac{a'_n x}{a_n} + \frac{b'_n - b_n}{a_n}\right) + b_n\right). \tag{5.8}$$

Valitaan x_1, siten että $H(x_1) \in (0, 1)$. Osoitetaan, että jono

$$\frac{a'_n x_1}{a_n} + \frac{b'_n - b_n}{a_n} \tag{5.9}$$

on rajoitettu. Oletetaan, että jonon 5.9 raja-arvo on ääretön. Koska F^n on kertymäfunktio, sen ominaisuuksien nojalla pätee, että

$$\lim_{n\to\infty} F^n\left(a_n\left(\frac{a'_n x_1}{a_n} + \frac{b'_n - b_n}{a_n}\right) + b_n\right) = 1.$$

Toisaalta nähdään myös, että

$$\lim_{n\to\infty} F^n\left(a_n\left(\frac{a'_n x_1}{a_n} + \frac{b'_n - b_n}{a_n}\right) + b_n\right) = \lim_{n\to\infty} F^n\left(a'_n x_1 + b'_n\right) = H(x_1) < 1,$$

mikä johtaa ristiriitaan. Jonon 5.9 täytyy olla siis ylhäältä rajoitettu.

Vastaavalla päättelyllä voidaan todeta, että jonon täytyy olla myös alhaalta rajoitettu. Jono on siis rajoitettu ylhäältä ja alhaalta. Lisäksi voidaan valita $x_2 \neq x_1$, siten että $H(x_2) \in (0,1)$, jota vastaava jono on rajoitettu. Tiedetään siis, että jonolle 5.9 on olemassa osajonot siten, että

$$\lim_{i\to\infty}\left(\frac{a'_{n_i}x_j}{a_{n_i}} + \frac{b'_{n_i} - b_{n_i}}{a_{n_i}}\right) = c_j, \quad j = 1, 2.$$

Näistä yhtälöistä voidaan todeta, että on olemassa raja-arvot osajonoille:

$$\lim_{i\to\infty}\frac{a'_{n_i}}{a_{n_i}} = a \quad \text{ja}$$

$$\lim_{i\to\infty}\frac{b'_{n_i} - b_{n_i}}{a_{n_i}} = b.$$

Valitaan x, siten että se on jakauman $H(x)$ mielivaltainen jatkuvuuspiste, ja lasketaan yhtälön 5.8 funktion raja-arvo osajonojen raja-arvojen avulla. Tällöin saadaan

$$H(x) = \lim_{i\to\infty} F^{n_i}\left(a_{n_i}\left(\frac{a'_{n_i}x}{a_{n_i}} + \frac{b'_{n_i} - b_{n_i}}{a_{n_i}}\right) + b_{n_i}\right) = G(ax+b), \tag{5.10}$$

jossa täytyy päteä $a > 0$, sillä $G(x)$ on kertymäfunktio. Väite 5.7 pätee nyt, sillä osajonojen raja-arvojen täytyy toteuttaa yhtälö $H(x) = G(ax+b)$. Oletetaan, että jonon 5.9 jollekin osajonolle pätee, että raja-arvot ovat

$$\lim_{i\to\infty}\frac{a'_{m_i}}{a_{m_i}} = a' \quad \text{ja}$$

$$\lim_{i\to\infty}\frac{b'_{m_i} - b_{m_i}}{a_{m_i}} = b'.$$

Tällöin voidaan yhtälön 5.10 päättelyllä perustella, että $H(x) = G(a'x+b') = G(ax+b)$. Nyrhisen luentomonisteen [15] lemman 2.1 nojalla $G(a'x+b') = G(ax+b)$ pätee jos ja vain jos $a' = a$ ja $b' = b$. Joten raja-arvot ovat yksikäsitteiset, ja lauseen ensimmäinen suunta on todistettu.

Todistetaan seuraavaksi toinen suunta väitteestä. Oletetaan, että 5.7 pätee. Lisäksi tiedetään, että $H(x) = G(ax+b)$ pätee. Tällöin voidaan siis sanoa, että

$$\lim_{n\to\infty} F^n\left(a_n\left(\frac{a'_n x}{a_n} + \frac{b'_n - b_n}{a_n}\right) + b_n\right) = G(ax+b).$$

Tästä yhtälöstä nähdään myös, että:

$$\lim_{n\to\infty} F^n\left(a_n\left(\frac{a'_n x}{a_n} + \frac{b'_n - b_n}{a_n}\right) + b_n\right) = \lim_{n\to\infty} F^n(a'_n x + b'_n) = H(x).$$

Täten lause on todistettu. □

Satunnaismuuttuja voi siis kuulua vain yhden jakauman ja sen affiinien muunnosten vaikutuspiiriin. Tämä tekee jakaumien luokittelusta helpompaa, sillä niitä voidaan luokitella sen mukaan, minkälaiseen vaikutuspiiriin jakauma kuuluu.

Näille vaikutuspiireille on olemassa yleisesti tunnettu lause, joka määrittelee kolme erilaista mahdollista jakaumatyyppiä, joihin satunnaismuuttuja voi kuulua. Lauseessa oletetaan, että raja-jakauma on ei-degeneroitunut.

Lause 5.4 (Fisher-Tippet-Gnedenko -teoreema)**.** *Jos satunnaismuuttuja X kuuluu jonkin ei-degeneroituneen jakauman $G(x)$ vaikutuspiiriin, niin jakauma $G(x)$ on jokin seuraavista ääriarvojakaumista:*

1. *Fréchet:* $\Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x^{-\alpha}), & x > 0, \end{cases}$
2. *Weibull:* $\Lambda(x) = \exp(-e^{-x})$,
3. *Gumbell:* $\Psi_\alpha(x) = \begin{cases} 0, & x \geq 0 \\ \exp(-|x|^{-\alpha}), & x < 0. \end{cases}$

Tätä teoreemaa ei todisteta tässä työssä, todistus löytyy esimerkiksi Resnickin kirjasta [18] sivulta yhdeksän. Teoreema on erittäin tärkeä ja se antaa selvän muodon mahdollisille maksimien jakaumille. Lause rajoittaa mahdolliset jakaumat kolmeen erilliseen jakaumaan, kunhan lauseen oletus on voimassa.

Nyt tarvittavat käsitteet Pickands-Balkema-de Haan -teoreemaan on esitelty, joten voidaan esitellä itse lause.

Lause 5.5 (Pickands-Balkema-de Haan -teoreema)**.** *Oletetaan, että satunnaismuuttujan X kertymäfunktio on $F(x)$. Kertymäfunktio $F(x)$ kuuluu johonkin ääriarvojakauman vaikutuspiiriin maksimin suhteen, jos ja vain jos on olemassa mitallinen funktio $\beta(u)$, jolle pätee:*

$$\lim_{u\to\infty} F_u(x) = \lim_{u\to\infty} \mathbb{P}(X - u \leq x | X > u) = G_{\alpha,\beta(u)}(x) = \begin{cases} 1 - \left(1 + \frac{x}{\alpha\beta(u)}\right)^{-\alpha} & , \alpha \neq 0 \\ 1 - e^{\frac{-x}{\beta(u)}} & , \alpha = 0, \end{cases}$$

jollakin $\alpha \in \mathbb{R}$.

Lausetta ei todisteta kokonaisuudessaan, vaan siitä käsitellään tutkimuksen kannalta olennaista tapausta. Tapauksessa tehdään oletus, että $F(x)$ kuuluu Fréchet-tyypin ääriarvojakauman vaikutuspiiriin ja todistetaan funktion $\beta(u)$ olemassaolo.

Todistus. Oletetaan, että kertymäfunktio $F(x)$ kuuluu Fréchet-ääriarvojakauman vaikutuspiiriin maksimin suhteen jollakin $\alpha \in \mathbb{R}$. Osoitetaan ensiksi, että $\bar{F}(x) \in \mathcal{R}_{-\alpha}$. Tiedetään, että

pätee

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \Phi_\alpha(x). \tag{5.11}$$

Valitaan x siten, että $\Phi_\alpha(x) > 0$, eli $x > 0$. Nyt huomataan yhtälöstä 5.11, että funktiolle $F(a_n x + b_n)$ pätee

$$\lim_{n\to\infty} F(a_n x + b_n) = 1, \tag{5.12}$$

sillä jos $\lim_{n\to\infty} F(a_n x + b_n) < 1$, niin pätisi

$$\lim_{n\to\infty} F^n(a_n x + b_n) = 0.$$

Joten selvästi yhtälö 5.12 pätee kaikilla $x > 0$. Koska yhtälön 5.11 kummankin puolen funktiot ovat jatkuvia, voidaan ottaa logaritmi puolittain, jolloin saadaan

$$\lim_{n\to\infty} n \log(F(a_n x + b_n)) = \log(\exp(-x^{-\alpha})). \tag{5.13}$$

Taylorin kehitelmän avulla voidaan laskea

$$\log(F(x)) = (F(x) - 1) - \frac{(F(x)-1)^2}{2} + \frac{(F(x)-1)^3}{3} - \dots$$

$$= -(1 - F(x))(1 - \frac{(F(x)-1)}{2} + \frac{(F(x)-1)^2}{3} - \dots) = -\bar{F}(x)(1 + o(1)),$$

missä $o(1) \to 0$, kun $F(x) \to 1$. Ja yhtälöstä 5.12 tiedetään, että $\lim_{n\to\infty} F(a_n x + b_n) = 1$. Joten voidaan sijoittaa tulos yhtälöön 5.13, jolloin saadaan

$$\lim_{n\to\infty} n \log(F(a_n x + b_n)) = \lim_{n\to\infty} -n\bar{F}(a_n x + b_n)(1 + o(1)) = -x^{-\alpha}.$$

Tämän yhtälön avulla voidaan perustella, että $\bar{F}(x) \in \mathcal{R}_{-\alpha}$. Tarkemmat yksityiskohdat jätetään tämän tutkimuksen ulkopuolelle, tarkempi esitys löytyy kirjasta [3] luvusta 8.14 tai Nyrhisen luentomonisteesta [15].

Karamatan esityslauseen seurauksen 5.1 nojalla, säännöllisesti vaihteleva funktio indeksillä $-\alpha$ voidaan esittää seuraavasti:

$$f_{-\alpha}(x) = c(x) \exp\left\{ -\int_z^x \frac{1}{p(t)} \mathrm{d}t \right\}. \tag{5.14}$$

Josta tiedetään, että $c(x) \to c \in (0, \infty)$ ja $\frac{t}{p(t)} \to \alpha$.

Nyt voidaan tutkia funktion $F_u(x)$ häntäfunktiota $\bar{F}_u(x) = 1 - F_u(x)$. Esitetään funktio $\bar{F}_u(x)$ jakauman häntäfunktion $\bar{F}(x)$ avulla. Lisäksi käytetään funktiota $p(u)$, joka on sama kuin yhtälössä 5.14 esiintyvä funktio. Voidaan laskea

$$\mathbb{P}\left(\frac{X-u}{p(u)} > x | X > u\right) = \bar{F}_u(xp(u)) = \frac{\bar{F}(u+xp(u))}{\bar{F}(u)}. \tag{5.15}$$

Tavoitteena on osoittaa, että häntäfunktio on yleistetyn Pareto-jakauman häntäfunktio. Koska häntäfunktiot ovat säännöllisesti vaihtelevia, voidaan sijoittaa yhtälön 5.14 tulos yhtälöön 5.15, jolloin saadaan

$$\begin{aligned}
\frac{\bar{F}(u+xp(u))}{\bar{F}(u)} &= \frac{c(u+xp(u))\exp\left\{-\int_z^{u+xp(u)} \frac{1}{p(t)}\mathrm{d}t\right\}}{c(u)\exp\left\{-\int_z^{u} \frac{1}{p(t)}\mathrm{d}t\right\}} \\
&= \frac{c(u+xp(u))}{c(u)}\exp\left\{-\int_z^{u+xp(u)} \frac{1}{p(t)}\mathrm{d}t + \int_z^{u} \frac{1}{p(t)}\mathrm{d}t\right\} \\
&= \frac{c(u+xp(u))}{c(u)}\exp\left\{-\int_u^{u+xp(u)} \frac{1}{p(t)}\mathrm{d}t\right\} \\
&= \frac{c(u+xp(u))}{c(u)}\exp\left\{-\int_u^{u+xp(u)} \frac{t}{p(t)}\frac{\mathrm{d}t}{t}\right\}.
\end{aligned}$$

Valitaan u riittävän iso, jotta voidaan merkitä $\frac{t}{p(t)} = \alpha + o(1)$, missä $o(1) \to 0$ kun $t \to \infty$. Tiedetään, että kun $u \to \infty$, myös $t \to \infty$. Vastaavasti integraalin ylärajassa merkitään: $p(u) = u(\alpha + o(1))^{-1}$. Saadaan tulokseksi:

$$\begin{aligned}
&(1+o(1))\exp\left\{-\alpha\int_u^{u+xu(\alpha+o(1))^{-1}} \frac{1}{t}\mathrm{d}t\right\} \\
&= (1+o(1))\exp\left\{-\alpha\left[\log\left(u+xu(\alpha+o(1))^{-1}\right) - \log(u)\right]\right\} \\
&= (1+o(1))\frac{(u(1+x(\alpha+o(1))^{-1}))^{-\alpha}}{u^{-\alpha}} = (1+o(1))(1+x(-\alpha+o(1))^{-1})^{-\alpha} \\
&\to \left(1+\frac{x}{\alpha}\right)^{-\alpha}, \quad u \to \infty.
\end{aligned}$$

Mikä todistaa lauseen, sillä nyt pätee, kun u on iso, että:

$$\mathbb{P}\left(\frac{X-u}{p(u)} \leq x | X > u\right) \sim G_{\alpha,1}.$$

Ja valitsemalla $\beta(u) = p(u)$, saadaan lauseen väitteessä oleva yleistetty Pareto-jakauma.

□

Tämä lause toimii teoreettisena perusteluna sille, että pyritään löytämään sopiva Pareto-jakauma. Visuaalisen tarkastelun pohjalta voimme olettaa, että potilaslaskujen jakauma kuuluu Fréchet-tyypin ääriarvojakauman vaikutuspiiriin maksimin suhteen, joten lauseen käyttö on perusteltua. Yleistetyllä ja tavallisella Pareto-jakaumalla on pieniä eroja, sillä tavallinen Pareto-jakauma on yleistetyn poikkeustapaus. Niiden käyttäytyminen on kuitenkin hyvin lähellä toisiaan, etenkin niiden indeksit vastaavat toisiaan, mikä on tämän tutkimuksen keskeisin tutkittava kohde. Seuraavaksi siirrytään vielä estimoimaan indeksiä käyttäen hyväksi lauseen tulosta.

5.2.3 Indeksin estimointi

Tarkastellaan tarkemmin ylityksen suuruuden tai paremmin tunnetulta nimeltään peaks-over-threshold -metodin ominaisuuksia indeksin estimoinnissa. Metodin suurin vahvuus on se, että empiiristä aineistoa käsitellessä, siinä voidaan keskittyä suurten kustannusten kannalta olennaisiin tapauksiin. Paksuhäntäisessä jakaumassa ei välttämättä ole niinkään väliä, miten alle keskiarvon olevat arvot ovat jakautuneet, vaan korkeampien kvantiilien suuruus on paljon kriittisempää. Varsinkin sellaisissa aineistoissa, joissa lukumäärällisesti pieni määrä suuria havaintoja aiheuttaa suurimman osan kustannuksista, on tärkeämpää pystyä ennustamaan suuria kustannuksia.

Yleisesti aineistoissa tämä lähestymistapa saattaa aiheuttaa ongelmia datan riittävyyden kannalta. Pienen todennäköisyyden tapahtumia esiintyy aineistoissa harvoin, ja mikäli aineisto on rajattu, niin tapahtumia saattaa olla vain muutamia. Tämä tekee estimaattoreista epätarkkoja ja tuloksista epäluotettavia. Tässä tutkimuksessa käsiteltävä potilaslaskujen aineisto on kuitenkin tarpeeksi laajaa, että siinä esiintyy useita pienen todennäköisyyden tapahtumia. Erityisesti suurien laskujen tutkiminen on tärkeää, sillä ne muodostavat ison osan kustannuksista.

Ylityksen suuruutta tutkittaessa käytetään ylityksen suuruuden kertymäfunktiota:

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u).$$

Voidaan määritellä seuraavaksi:

$$\bar{F}(u + x) = \bar{F}(u)\bar{F}_u(x).$$

Josta nähdään, että raja-arvon u ylittävien tapahtumien todennäköisyyttä voidaan arvioida estimoimalla erikseen funktioita $\bar{F}(u)$ ja $\bar{F}_u(x)$. Ensimmäiseen voidaan soveltaa empiiristä kertymäfunktioita, ja toiseen voidaan juuri todistetun Pickands-Balkema-de Haan -teoreeman mukaan käyttää yleistettyä Pareto-jakaumaa. Yleinen estimaatti saadaan Smithin esittämästä yhtälöstä [22],

$$\hat{\bar{F}}(u + x) = \frac{N(u)}{n}\left(1 + \frac{y}{\hat{\beta}\hat{\alpha}}\right)^{-\hat{\alpha}}.$$

Missä n on koko aineiston havaintojen määrä, ja N on rajan u ylittävien havaintojen määrä. Tämän estimaattorin asymptoottiset ominaisuudet ovat paremmat kuin Hillin estimaatin, sillä

se toimii myös silloin kun satunnaismuuttujan kertymäfunktio ei ole tarkasti Pareto, vaan riittää, että se on muotoa $\bar{F}(x) \approx x^{-\alpha}L(x)$, missä $L(x)$ on hitaasti vaihteleva funktio.

Seuraavaksi halutaan tarkastella varsinaista hännän indeksiä, eli löytää α, jolla $\bar{F}_u(x) \in \mathbb{R}_{-\alpha}$. On olemassa monia erilaisia tapoja estimoida tätä indeksiä, sillä yleistettyä Pareto-jakaumaa on tutkittu paljon. Pickands-Balkema-de Haan -teoreeman nojalla on perusteltua olettaa jakauman $F_u(x)$ olevan yleistetyn Pareto-jakauman mukainen, ja siten näiden estimaattoreiden käyttö on perusteltu. Colesin ja Stuartin artikkelissa [4] on esitelty näistä suurin osa. Myös Pickands esittelee artikkelissan [16] yhden estimaattorin indeksille, jota kutsutaan Pickandsin estimaattoriksi.

Pickandsin estimaattorissa käytetään järjestettyä jonoa havaituista satunnaismuuttujista $X_1, \ldots, X_n$, joita merkitään $Z_1, Z_2, \ldots, Z_n$, missä Z_1 on suurin havainto. Estimaattiin tulee valita luku M, josta oletetaan, että $4M$ isointa havaintoa sisältävät kaiken oleellisen tiedon jakauman hännästä. Luvun M tulee olla paljon pienempi kuin n, jotta tutkiminen olisi mielekästä. Kun M on valittu, saadaan estimaatti indeksille α seuraavalla kaavalla:

$$\hat{\alpha}^{-1} = \frac{1}{\log(2)} \log\left(\frac{Z_M - Z_{2M}}{Z_{2M} - Z_{4M}}\right).$$

Pareto-jakauman momenttien avulla, mikäli ne ovat olemassa, voidaan myös estimoida indeksiä. Menetelmä on esitelty artikkelissa [9]. Tässä menetelmässä käytetään hyväksi yhtälöä:

$$\mathbb{E}\left(1 + \frac{X}{\alpha\beta}\right)^r = \frac{1}{(1 - \frac{r}{\alpha})}, \quad \text{jos} \, 1 - \frac{r}{\alpha} > 0.$$

Josta saadaan estimaattori indeksille:

$$\hat{\alpha} = \frac{1}{2}\bar{x}\left(\frac{\bar{x}^2}{\bar{s}^2} + 1\right).$$

Momenttiestimaattorilla on hyvät asymptoottiset ominaisuudet, joten sen käyttäminen on hyvin perusteltua.

Viimeinen esiteltävä estimaattori on todennäköisyyksillä painotettujen momenttien -metodi, englanniksi "Probability-weighted moments", lyhyemmin PWM, joka on esitelty tarkemmin artikkelissa Colesin artikkelissa [4]. Tämä metodi on hyvin lähellä momenttien perusteella tehtyä estimointia, mutta se antaa enemmän painoa hännän arvoille. Tutkitaan todennäköisyydellä painotettuja momentteja:

$$\beta_r = \mathbb{E}[X(F(X))^r].$$

Jota voidaan estimoida seuraavalla kaavalla:

$$\hat{\beta}_r = \frac{1}{n}\sum_{i=1}^{n} x_i(\tilde{F}(x_i))^r,$$

jossa $\tilde{F}$ on empiirinen kertymäfunktio. Tätä suuretta voidaan verrata yleistetyn Pareto-jakauman teoreettisiin arvoihin, ja ratkaista yhtälöryhmästä numeerisesti jakauman parametrien arvot.

Kaikissa näissä estimaatiomenetelmissä raja-arvon u valinta vaikuttaa lopputulokseen, sillä se vaikuttaa käytettävien havaintojen lukumäärään ja suuruuteen. Raja-arvon valinta on yleensä ottaen subjektiivinen päätös, vaikka teoreettisesti valinnan pystyisi tekemään asymptoottisesti tehokkaasti, niin käytännössä tämä on yleensä mahdotonta. Raja-arvon valinnassa joudutaan tasapainottelemaan kahden asian välillä: tarpeeksi suuren otoskoon ja mallioletusten täyttymisen. Jos raja-arvo on erittäin suuri verrattuna aineistoon, niin tutkittavia havaintoja on vähän, mutta jakauma on todennäköisemmin lähellä yleistettyä Pareto-jakaumaa. Toisaalta jos raja-arvo on pieni, otoskoko on suurempi, mutta jakauma ei välttämättä ole lähellä Pareto-jakaumaa. Lisäksi empiirisessä tutkimuksessa tulee ottaa huomioon esimerkiksi mahdolliset mittausvirheet, jotka voivat varsinkin suurien havaintojen kohdalla vaikuttaa lopputulokseen huomattavasti.

Embrecthsin, Klüppelbergin ja Mikoschin kirjassa [6], osiossa 6.5, kerrotaan tarkemmin raja-arvon valinnasta. Yksi vaihtoehto on tutkia milloin odotettu ylitysfunktio on lineaarinen, ja valita raja-arvo sen mukaan, sillä ylitysfunktion lineaarisuus implikoi jakauman olevan Pareto.

Esitellään seuraavaksi eri raja-arvon u valinnoilla, ja eri menetelmillä laskettuja estimaatteja laskujen jakauman indeksistä. Raja-arvot on valittu seuraavasti:

1. $u_1 = 2000$, raja-arvo jossa odotetun ylityksen funktio on lineaarinen.
2. $u_2 = 8660$, 90-kvantiili laskuista pyöristettynä.
3. $u_3 = 14000$, 95-kvantiili laskuista pyöristettynä.
4. $u_4 = 33000$, 99-kvantiili laskuista pyöristettynä.
5. $u_5 = 63000$, valittu 400 suurinta laskua.

Estimaatit ovat esiteltyinä taulukossa 5.3. Niiden laskemiseen käytettiin R:n POT-pakettia, jonka käyttö on esitelty artikkelissa [19]. Taulukosta 5.3 huomataan, että eri metodeilla saadaan erilaiset estimaatit indeksin arvoille. Kaikki estimaatit vaikuttavat kuitenkin olevan lähellä arvoa 2.3. Pickandsin estimaattorin tulos vaihtelee huomattavasti eri raja-arvoilla, mutta momentti- ja PWM-estimaattori pysyvät suhteellisen stabiileina.

Kokonaisuudessaan taulukon 5.3 pohjalta voidaan sanoa, että jakauman indeksi α voisi olla välillä $\alpha \in (2.2, 2.4)$. Tästä tulee esiin paksuhäntäisten jakaumien tutkimisen epätarkkuus. Koska isoimmat arvot vaikuttavat merkittävästi indeksin estimaattiin, ja niiden lukumäärä on yleensä vähäinen, niin estimointi saattaa olla epätarkkaa. Mutta tämän taulukon, ja aiemman visuaalisen tarkastelun pohjalta voidaan perustellusti sanoa, että häntäfunktion indeksille α

Taulukko 5.3: Taulukko ylityksen jakauman indeksin estimaateista eri metodeilla ja eri raja-arvon u arvoilla. Metodit ovat: Pickandsin estimaattori, momenttiestimaattori ja todennäköisyyksillä painotettu momenttiestimaattori.

Metodi	u=2000	u=8660	u= 14000	u=33000	u=63000
Pickands	2.76	2.33	2.22	2.75	1.52
Momentit	2.36	2.37	2.37	2.37	2.42
PWM	2.37	2.33	2.34	2.24	2.16

voisi päteä, että $\alpha \approx 2.3$. Tämä poikkeaa hieman visuaalisesta tarkastelusta, jossa hahmoteltiin, että Pareto-jakauma indeksillä 2.1 sopisi hyvin jakaumaan. Indeksi $\alpha \approx 2.3$ vaikuttaa uskottavalta ratkaisulta häntäfunktion indeksiksi.

Jakauman indeksi 2.3 tarkoittaisi sitä, että jakauman odotusarvo ja varianssi ovat olemassa, mutta vinoutta ja huipukkuutta ei ole määritelty. Tämä vaikuttaa uskottavalta tulokselta jakauman kannalta.

Tarkemmalla analyysillä voitaisiin vielä estimoida yleisen Pareto-jakauman muita parametrejä, ja sitten iteratiivisesti estimoida uudestaan indeksin α arvoa. Tätä prosessia voidaan toistaa useamman kerran ja saada uudet estimaatit parametreille. On myös mahdollista, että iterointimenetelmällä parametrit eivät tarkennu mihinkään tiettyyn arvoon. Tällaisella menetelmällä voitaisiin tutkia vielä tarkemmin laskujen jakaumaa, mutta se ei kuulu tämän tutkimuksen piiriin.

6 Johtopäätökset

Tutkimus aloitettiin tarkastelemalla sairaanhoitopiirin kokonaiskustannuksia, muodostamalla ensiksi lineaariset regressiomallit eri kustannustyypeille. Sen jälkeen muodostettiin VAR(p)-malli kokonaiskustannuksille ja ennustettiin kustannusten käyttäytymistä tulevaisuudessa. Tutkimuksessa käytetty data on kerätty vuosilta 2012-2019, ja ennuste on luotu vuosille 2020-2022, mutta koronavirustilanteen takia ennuste vuodelle 2020 ei todennäköisesti ole tarkka. Tilanne saattaa myös vaikuttaa tulevien vuosien kustannuksiin, mutta tarkkoja arvioita on mahdotonta tehdä tässä vaiheessa.

Aineistoa tutkittaessa huomattiin, että yksi selvästi kustannusten kasvuun vaikuttava tekijä on se, että päivystyksellisen hoidon osuus kustannuksista ja hoidon määrästä on kasvanut. Tämä voi johtua osittain sairaanhoitopiirin hoitokäytännöistä, joissa pyritään mahdollistamaan potilaan aikainen kotiutuminen. Myös väestönkasvu ja ikääntyminen selittää osaltaan kustannusten kasvua. Suurempi väestömäärä tarkoittaa suurempaa hoidontarvetta, ja vanhemmat ihmiset tarvitsevat myös yleensä ottaen enemmän hoitoa kuin nuoremmat.

Henkilöstökustannusten osuus sairaanhoitopiirin kokonaiskustannuksista on merkittävä, se on selvästi suurin kustannuslaji, joka on myös kasvanut tutkitulla ajanjaksolla. Osittain kasvu johtuu todennäköisesti inflaatiosta ja suuremmasta hoidontarpeesta. Vaikka sairaanhoitopiirissä tuotettujen hoitopäivien lukumäärä on pysynyt tasaisena, niin poliklinikkakäyntien lukumäärä on kasvanut selvästi. Tämä tarkoittaa suurentunutta henkilöstön tarvetta, sillä poliklinikkakäynnit vaativat henkilöstöä.

Tutkimuksen viimeisessä osiossa tarkasteltiin yksittäisten potilaiden laskujen jakaumaa. Lopputuloksena todettiin, että laskujen jakauma on todennäköisesti paksuhäntäinen, mikä tarkoittaa, että on olemassa suhteellisen suuri todennäköisyys sille, että sairaanhoitopiiriin tulee kalliita potilaita. Tällaisiin tilanteisiin on hyvä varautua, sillä kalliit potilaat saattavat aiheuttaa merkittävän kustannuspiikin sairaanhoitopiirille. Tutkimuksessa estimoitiin myös sitä, että kuinka paksuhäntäinen jakauma on, ja yhdeksi vaihtoehdoksi saatiin, että jakauman indeksi saattaisi olla lähellä arvoa 2.3. Tämä vastaa suunnilleen sitä, että jakauman odotusarvoa 25 kertaa suuremman havainnon todennäköisyys on 0.1%.

Tällaiset yksittäiset potilastapaukset saattavat tulla erittäin kalliiksi sairaanhoitopiirille, varsinkin jos niitä sattuu useampi lyhyen ajan sisällä. Tällaisiin kustannuksiin varautuminen voi olla haastavaa, mutta riskien ennakoiminen on oleellinen osa kustannusten hallintaa, ja se tulisi ottaa huomioon budjetteja suunniteltaessa. Yksi mahdollisuus varautumiseen olisi ottaa vakuutus kalliiden potilaiden varalta.

Viitteet

[1] H. Akaike. "Information Theory And An Extension Of The Maximum Likelihood Principle". *Proceedings of the 2nd International Symposium on Information Theory* (1993). Toim. F Csaki BN Petrov, s. 267–281.

[2] A. A. Balkema ja L. de Haan. "Residual Life Time at Great Age". eng. *The Annals of Probability* 2.5 (1974), s. 792–804. DOI: 10.1214/aop/1176996548.

[3] N. H. Bingham, C. M. Goldie ja Jef L. Teugels. *Regular Variation*. eng. Cambridge University Press, 1987. DOI: 10.1017/CBO9780511721434.

[4] S. Coles ja M. J. Dixon. "Likelihood-Based Inference for Extreme Value Models". eng. *Extremes* 2.1 (1999), s. 5–23. DOI: 10.1023/A:1009905222644.

[5] R. Durrett. "Probability: Theory and Examples". Teoksessa: 5. painos. Cambridge: Cambridge University Press, 2019. ISBN: 9781108473682.

[6] P Embrechts, C. Klüppelberg ja T. Mikosch. *Modelling extremal events : for insurance and finance*. Applications of mathematics. Berlin: Springer, 1997, s. 645.

[7] S. Ghosh ja S. I. Resnick. "A Discussion on Mean Excess Plots". *Stochastic Processes and their Applications* 120 (heinäkuu 2009). DOI: 10.1016/j.spa.2010.04.002.

[8] B. M. Hill. "A Simple General Approach to Inference About the Tail of a Distribution". eng. *The Annals of Statistics* 3.5 (1975), s. 1163–1174.

[9] J. R. M. Hosking ja J. R Wallis. "Parameter and Quantile Estimation for the Generalized Pareto Distribution". eng. *Technometrics* 29.3 (1987), s. 339–349. DOI: 10.1080/00401706.1987.10488243.

[10] C. M. Jarque ja A. K. Bera. "Efficient tests for normality, homoscedasticity and serial independence of regression residuals". eng. *Economics Letters* 6.3 (1980), s. 255–259. DOI: 10.1016/0165-1765(80)90024-5.

[11] T. Kilpeläinen. *Mitta- ja integraaliteoria*. 2004.

[12] Terveyden ja hyvinvoinnin laitos. *Sotkanet.fi*. 2020 (Haettu 25. tammikuuta). URL: https://sotkanet.fi/sotkanet/fi/haku?g=219.

[13] H. Lütkepohl. *Introduction to multiple time series analysis*. Berlin: Springer-Verlag, 1991.

[14] H. Lütkepohl ja M. Krätzig. *Applied Time Series Econometrics*. Cambridge: Cambridge University Press, 2004.

[15] H. Nyrhinen. *Äärimmäisten ilmiöiden teoriaa*. Helsingin yliopisto, 2016.

[16] J. Pickands. "Statistical Inference Using Extreme Order Statistics". eng. *The Annals of Statistics* 3.1 (1975), s. 119–131. DOI: 10.1214/aos/1176343003.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/.

[18] S. I. Resnick. *Extreme values, regular variation and point processes*. New York, NY: Springer, 1987.

[19] M. A. Ribatet. *A User's Guide to the POT Package (Version 1.0)*. 2006. URL: http://cran.r-project.org/.

[20] P. Saikkonen. *Moniulotteiset aikasarjat*. 2010.

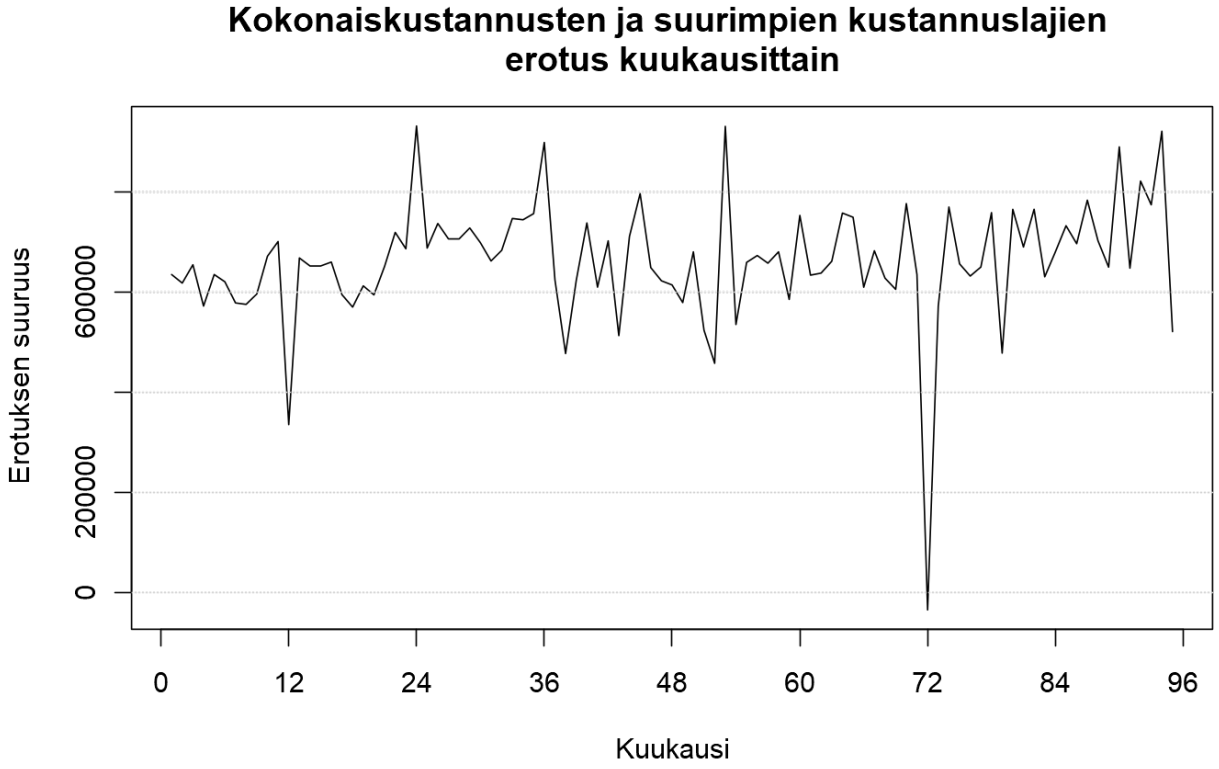
[21] G. Schwarz. "Estimating the Dimension of a Model". eng. *The Annals of Statistics* 6.2 (1978), s. 461–464. DOI: 10.1214/aos/1176344136.

[22] R. L. Smith. "Estimating Tails of Probability Distributions". eng. *The Annals of Statistics* 15.3 (1987), s. 1174–1207. DOI: 10.1214/aos/1176350499.

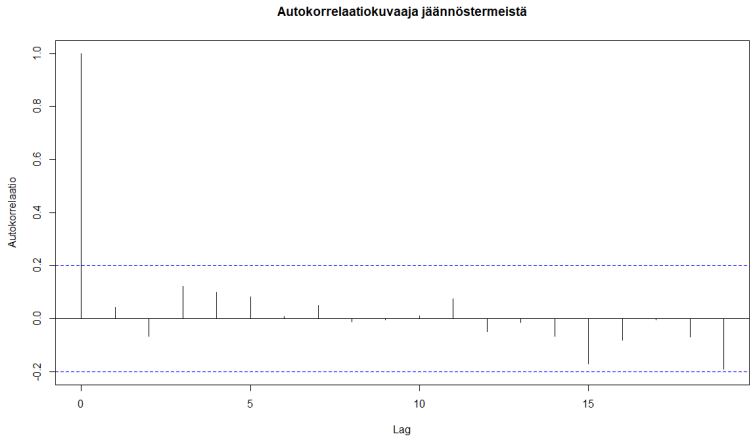
[23] S. Urbanek ja J. Horner. *Cairo: R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) and Display (X11 and Win32) Output*. R package version 1.5-12. 2020. URL: https://CRAN.R-project.org/package=Cairo.

A Liitteet

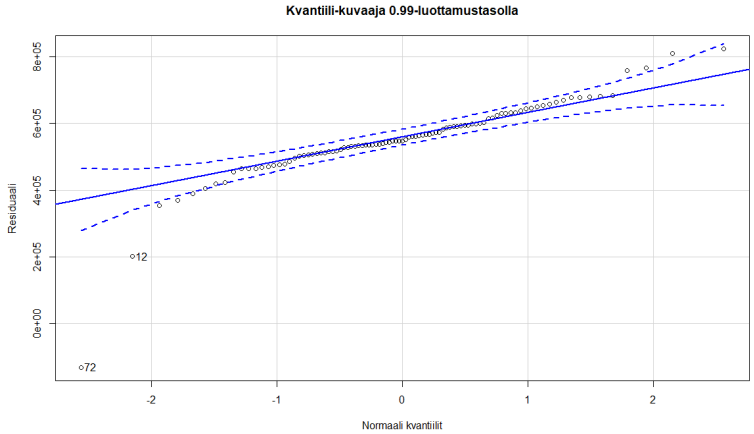
Liitteissä on esitelty kuvaajia, jotka ovat tutkimuksen kannalta oleellisia, mutta eivät ole lukemisen sujuvuuden kannalta välttämättömiä.



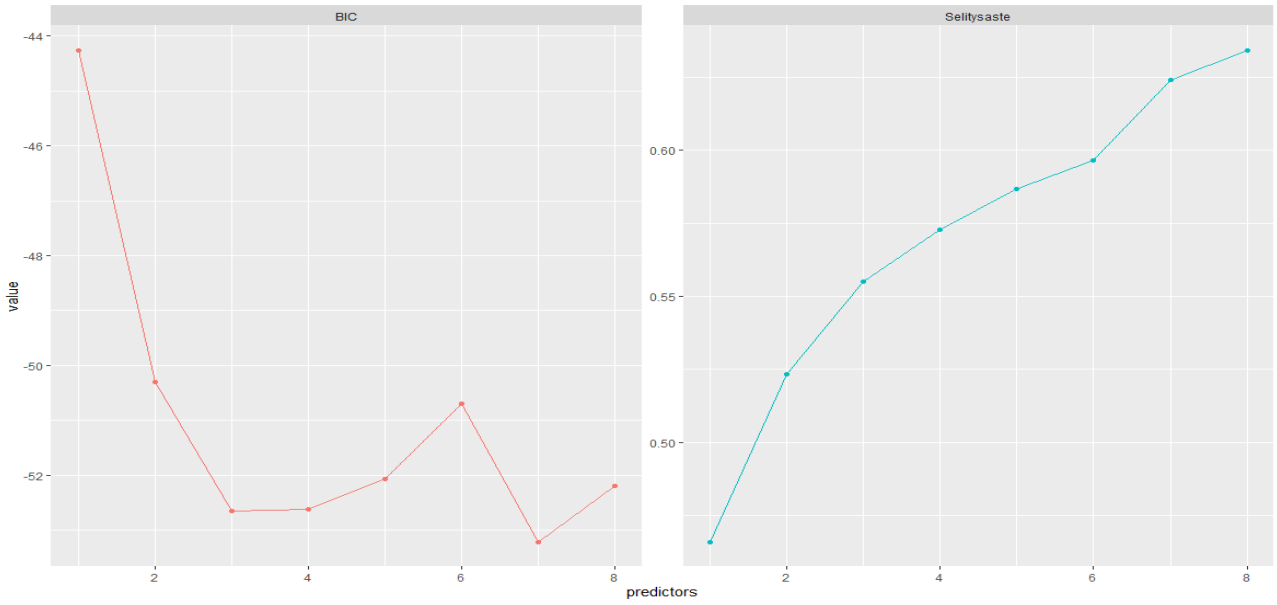
Kuva A.1: Kokonaiskustannusten ja seitsemän suurimmankustannuslajin erotukset kuukausitasolla. Yleisesti erotus näyttäisi pysyvän säännöllisellä tasolla. Kuukauden 72, eli vuoden 2017 joulukuussa on iso poikkeama.



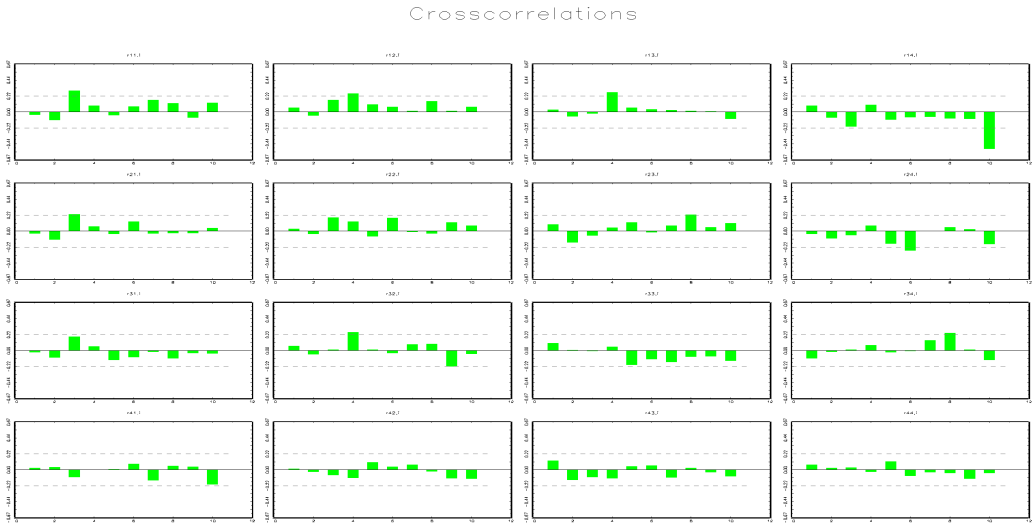
Kuva A.2: Kokonaiskustannusten ja seitsemän suurimman kustannuslajin erotuksen jäännöstermien autokorrelaatio



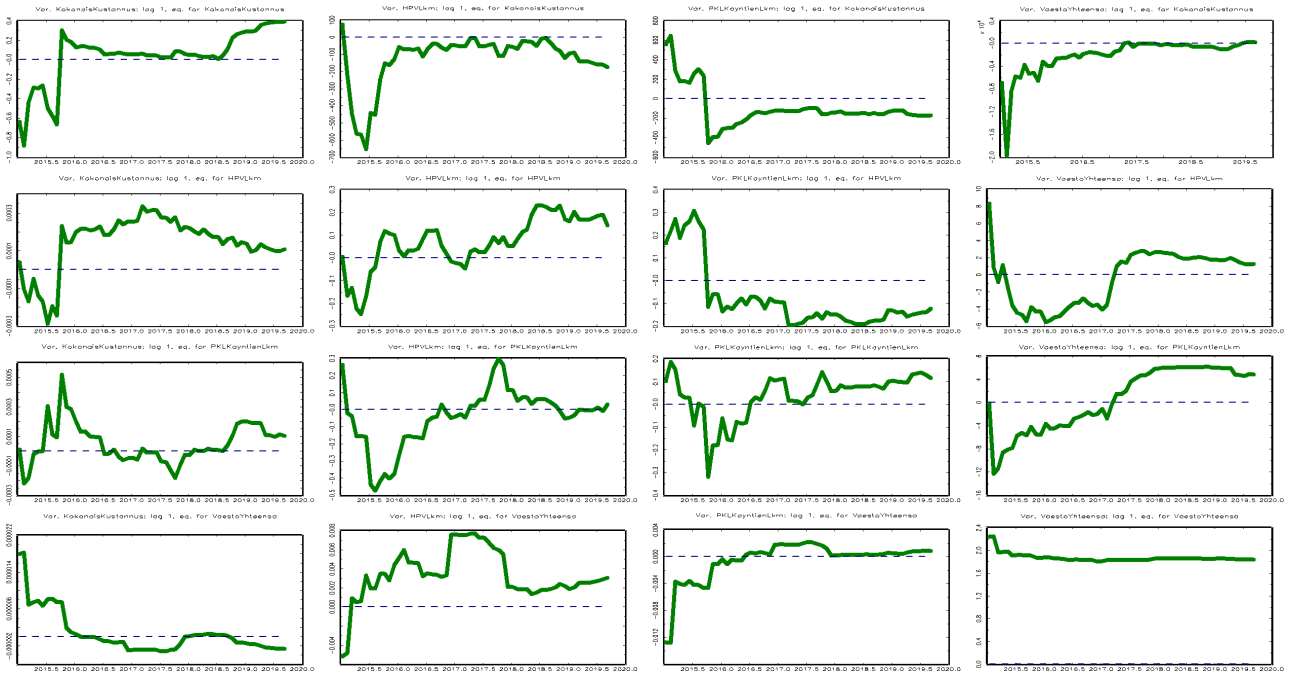
Kuva A.3: Kokonaiskustannusten ja seitsemän suurimman kustannuslajin erotuksen jäännöstermien kvantiilit verrattuna normaalijakaumaan



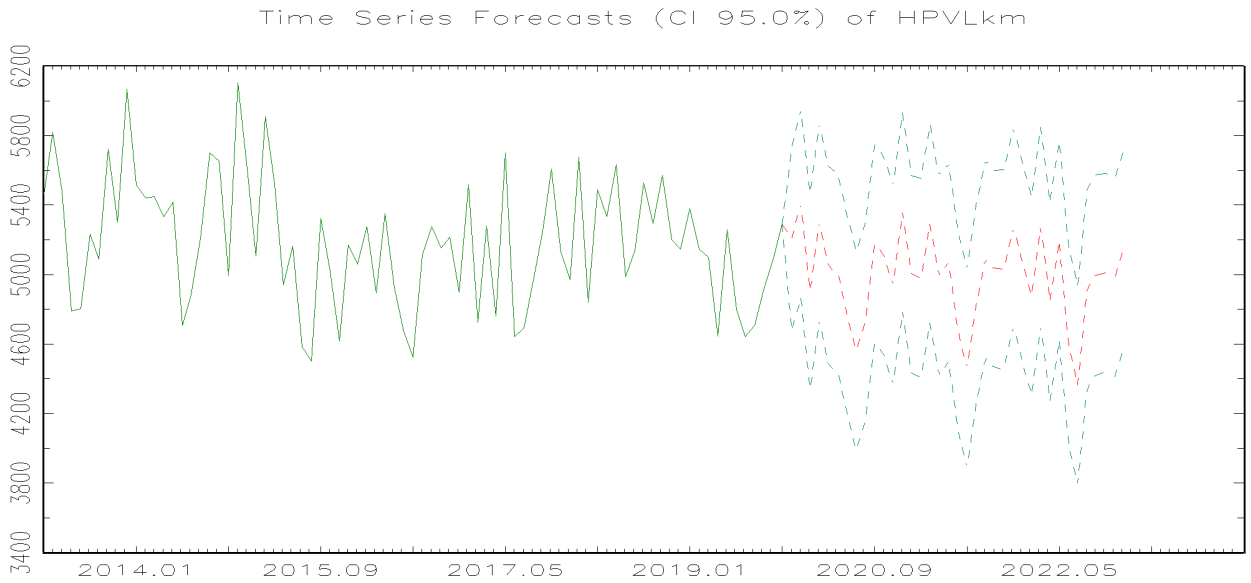
Kuva A.4: Kuvaaja informaatiokriteerin arvoista ja selitysasteesta.



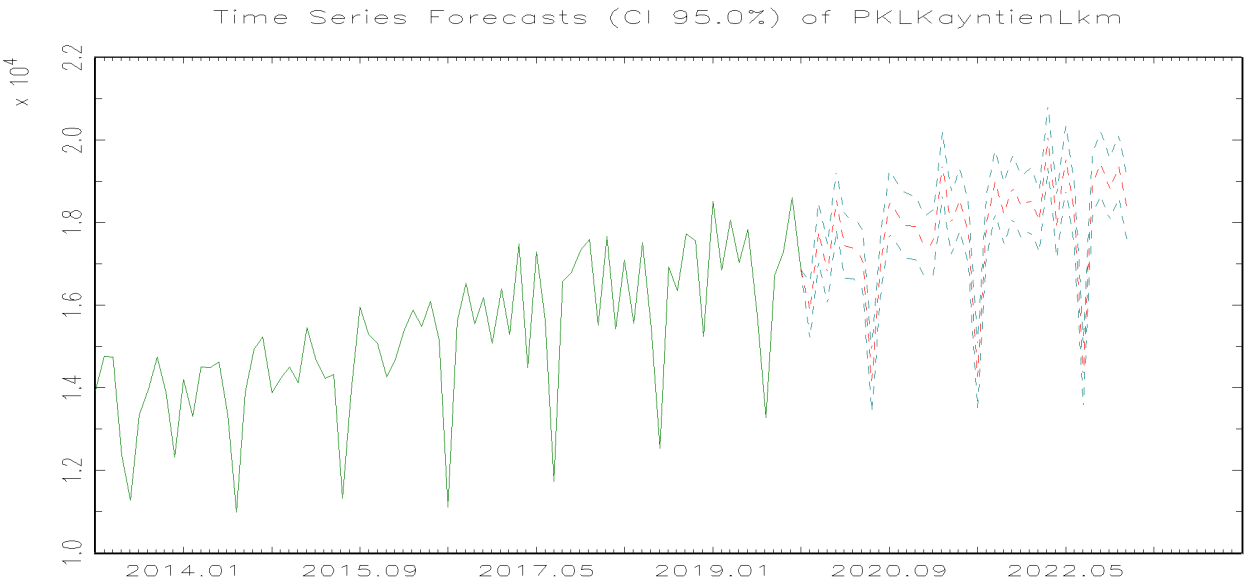
Kuva A.5: VAR-mallin residuaalien ristikorrelaatiot



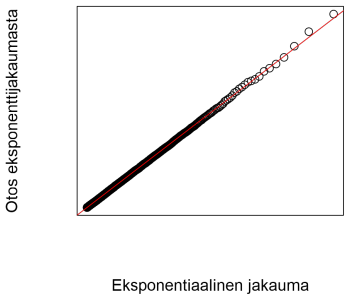
Kuva A.6: VAR-mallin rekursiiviset kertoimet



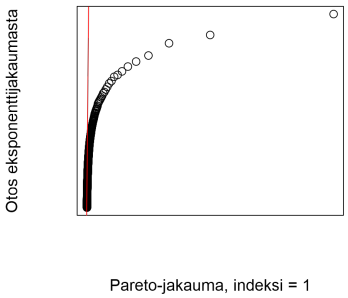
Kuva A.7: Hoitopäivien lukumäärä



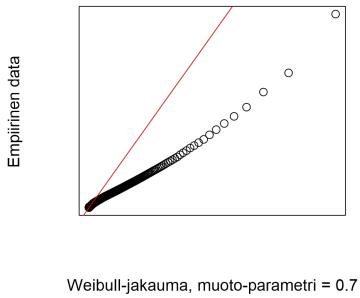
Kuva A.8: Poliklinikkakäyntien lukumäärä



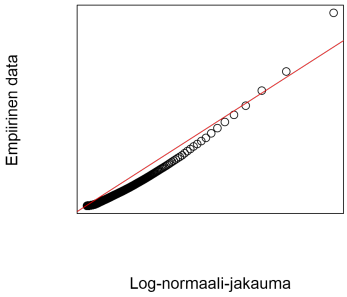
(a) Eksponenttijakauma verrattuna itseensä



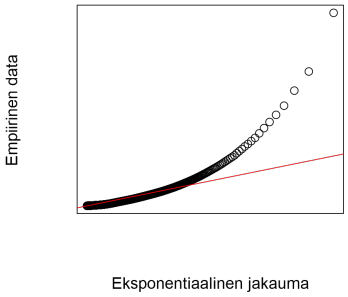
(b) Eksponenttijakauma verrattuna Pareto-jakaumaan



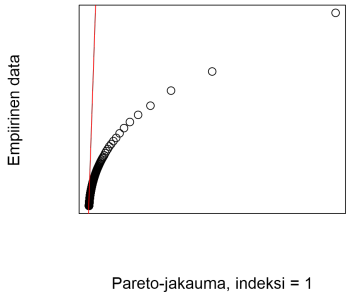
(c) Laskujen kvantiilit verrattuna Weibull-jakaumaan



(d) Laskujen kvantiilit verrattuna log-normaalijakaumaan



(e) Laskujen kvantiilit verrattuna eksponenttijakauman kvantiileihin



(f) Otos laskuista verrattuna Pareto-jakaumaan

Kuva A.9: QQ-kuvaajia, joissa verrataan sekä laskuja, että otosta eksponenttijakaumasta erilaisiin jakaumiin. Kuviin on piirretty $x = y$ suora havainnollistamaan pisteiden asettumista.